Quantum Physics of Semiconductor Materials and Devices

Part I

Fundamentals

1

And off We Go!

1.1 Beyond belief 3
1.2 A brief history of semiconductors 4
1.3 Future 6
1.4 These boots are made for walking 7
1.5 Chapter summary section 8
Further reading 8
Exercises 9

In this chapter, we get acquainted with the reason, the structure, and the best way to use the book. Each chapter will begin with a boxed summary of this form:

- Why this book?
- What is in this book?
- How can you make the best use of this book?

The summary is aimed to start out with a well-defined set of questions that are answered in the chapter.

1.1 Beyond belief

How did the smallest and the lightest fundamental particle in nature with mass –the **electron**– enable me to type these words on a computer? How did the electron help light up the screen of my laptop? How did it help remember all of this text as the battery of my laptop ran out? And how did it help kick those photons that power emails, and download this file and read these words? Explaining this remarkable story is a lion's share of what we will do in this book.

The invention of the steam engine made a mockery of what was believed to be the limit of the speed at which heavy mechanical objects could be moved. In much the same way, the discovery of the electron, and in particular the discovery of the class of materials called semiconductors has made a mockery of what was believed possible for three of the deepest and most profound human endeavors:

- **computation:** performing logic to produce information,
- **memory:** storing information, and
- **communication:** transmitting and receiving information.

Our understanding of the inner workings of electrons in semiconductors has given us powers beyond belief. We depend on semiconductors today to see and talk with family and/or friends on the other side of the planet. Semiconductors empower us to generate energy from sunlight, predict weather, diagnose and treat diseases, decode the DNA, design and discover new drugs, and guide our cars in the streets as deftly as satellites in deep space. They have placed the entire recorded history of information about the universe in the palm of our hands.

In this book, we will dig in to understand how all of this is possible. The semiconductor revolution is one of the most remarkable human adventures in the entire recorded history of science. The revolution has been powered by the combined efforts of scientists and engineers from several fields. The contributions of mathematicians and physicists, of chemists and materials scientists, and electrical engineers have made this possible. We all acknowledge that the artificial labeling of our skills serves an administrative purpose[1], and has no basis in science! Nowhere is this highlighted more than in the field of semiconductors in which "engineers" have been awarded Nobel Prizes in Physics, and physicists and chemists have founded successful "engineering" semiconductor companies such as Intel on their way to becoming billionaires.

1.2 A brief history of semiconductors

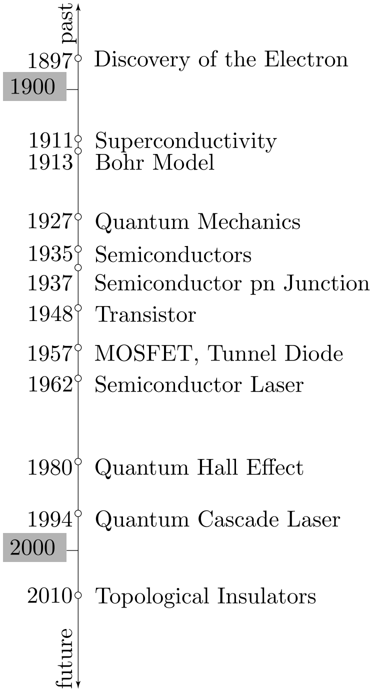


Fig. 1.1 A brief history of electronic materials.

Fig. 1.2 J. J. Thomson discovered the electron in 1897 at the Cavendish Laboratory (see Fig. 25.3 too). Awarded the 1906 Nobel Prize in Physics. Seven of his students went on to win Nobel Prizes.

Fig. 1.1 shows a broad timeline of topics on the physics and applications of semiconductors that are relevant in this book. Insulators and metals were known centuries ago. Their properties were studied, and theories were developed to explain them. In Chapter 2, we will intercept this emerging story during the period right after the discovery of the electron in 1897 by J. J. Thomson (Fig. 1.2). The concept of the electron subject to classical laws of mechanics, electromagnetism, and thermodynamics was used by Paul Drude to explain many properties of metals that had remained mysterious till that point of time.

It comes as a surprise to many today that superconductivity was experimentally discovered in 1911, *decades* before semiconductors were recognized as a distinct electronic state of matter. But our understanding of the physics of semiconductors developed much more rapidly than that of superconductors, as did their applications – as highlighted in Fig. 1.1. This book will make the reasons clear why this is so – that electrons in semiconductors are at first glance "simpler" to understand because they act independently as *single particles*, whereas electrons in superconductors pair up and are strongly *correlated*. However, the simplicity of electrons in semiconductors was just a mask. Applications following the invention of the semiconductor transistor in 1948 led to a very high level of control, perfection, and understanding of semiconductors, which has led to the discovery of several layers of richer physical behavior.

Esaki discovered electron tunneling in a solid – a genuine quantum mechanical effect highlighting the wave-nature of electrons – in semiconductor p-n diodes in 1957, around the same time Bardeen, Cooper, and Schrieffer proposed the eponymous BCS theory of superconductivity. The semiconductor transistor miniaturized the vaccum tube amplifier and made it far more compact, rugged, and energy efficient, setting up the stage for its use for digital logic and memory, enabling computation. In much the same way, by understanding how electrons in semiconductors interact with photons, the invention of the semiconductor laser diode in the 1960s and 1970s shrunk the solid-state laser and revolutionized photonics, enabling semiconductor lighting and optical communications.

---

[1] Harvard's "Division of Engineering and Applied Sciences" or DEAS changed its name to SEAS, the "School of Engineering and Applied Sciences". No one wants to create a division between applied sciences and engineering indeed!

In 1980, von Klitzing discovered the integer quantum Hall effect while investigating electron transport in the silicon transistor at low temperatures and at high magnetic fields. This was followed by the discovery of the fractional quantum Hall effect in 1982 in high-quality quantum semiconductor heterostructures grown by molecular beam epitaxy (MBE) by Tsui, Stormer, and Gossard. The quantum Hall effects revealed a whole new world of condensed matter physics, because under specific constraints of lower dimensionality, and the right electric and magnetic fields, the behavior of electrons defied classification into any of the previously identified phases: metals, insulators, semiconductors, or superconductors.

The effort to classify the quantum Hall state has led to the unveiling of new electronic and magnetic phases of electrons in solids: based on the *topology* of their allowed energy bands. This line of thought has led to the discovery of *topological insulators*, the newest addition that significantly enriches the field of electronic phases of matter. The mathematical language of topology was developed in the early part of the 20th century. Slowly but surely, the deep connection of several physical phenomena such as electronic polarization, magnetization, and spin-Hall effects to the topological and geometric aspects of their quantum states are being discovered.

The discovery of high-temperature layered cuprate superconductors in 1987 by Bednorz and Muller again rocked the field of superconductivity, and the field of condensed matter physics. That is because the tried-and-tested BCS theory of superconductivity is unable to explain their behavior. At the forefront of experiments and theory today, there is a delightful connection being established between topological insulators (that have emerged from semiconductors), and superconductivity (that have emerged from metals). There are now proposals for topological superconductivity, and semiconductor–superconductor heterostructures that can make electrons behave in much stranger ways – in which they pair up to lose their charge and spin identities and can "shield" themselves from electric and magnetic fields. Such particle-antiparticle pairs, called Majorana fermions can enable robust quantum-bits (qubits) for quantum computation in the future. If all this sounds like science fiction[2] – let me assure you it is not, the race is on to experimentally find these elusive avatars of the electron in many laboratories around the world!

[2] The 2009 movie Avatar showed humans trying to mine the fictional material "Unobtanium" from Pandora – a room-temperature superconductor.

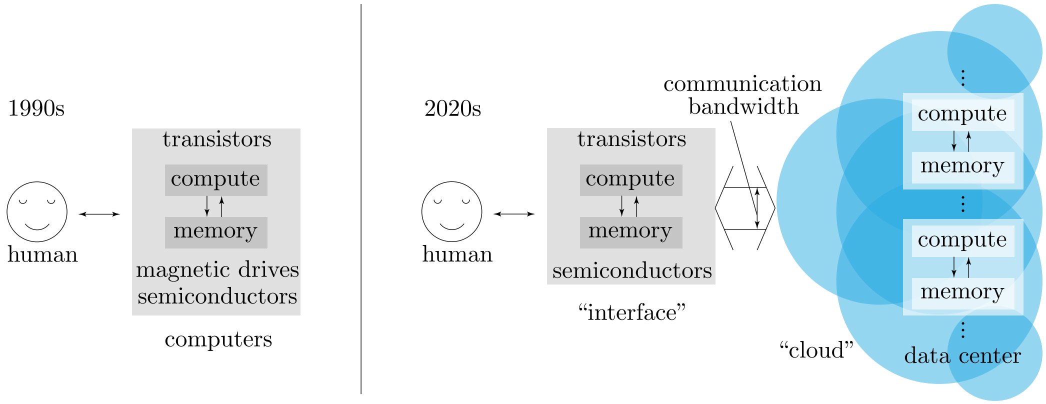

1990s
transistors
compute
memory
human
magnetic drives
semiconductors
computers
2020s
communication bandwidth
transistors
compute
memory
human
semiconductors
"interface"
compute
memory
compute
memory
"cloud"
data center

Fig. 1.3 How semiconductor technologies have changed the handling of data and information over the last two decades.

1.3 Future

Semiconductors today power the "information age" in which data (or information) is produced in high volumes, and transported at the speed of light. Fig. 1.3 shows the underlying structure of information systems. In the 1990s, most computers were stand-alone devices that we interacted with and programmed directly. They boasted semiconductor integrated circuit-based microprocessors for computation and logic, semiconductor based random-access memories (RAMs), and magnetic (or spin-based) hard drives or read-only memories (ROMs) as two types of memory. The communication of data was achieved through wires directly connected to the computers. Today, the wires are slowly vanishing as they are replaced by *photons* – both of the microwave or RF kind that propagate through the atmosphere, and the infrared kind that propagate in optical fibers, shuttling massive amounts of data at the speed of light. A large and increasing part of the computing infrastructure resides in the "cloud". The cloud is loosely made of a large and distributed network of semiconductor microprocessors and memories that may be housed in server farms. Much of the heavy computational tasks and memory storage is thus not performed on the computers in front of us with which we directly interact. The communication links that connect our interface devices (or so-called "Edge-Devices") with the cloud are therefore increasingly important, and their bandwidth will determine the efficiency of such networked information systems.

Just as a machine as complicated as a car needs to be driven, the information system needs instructions in the form of computer and system "programs", or "code". These instruction sets, no different in principle than sets of instructions to drive a car, set the whole system in motion and control its behavior, just as a driver weaving through traffic. The computational and memory capacity created by miniaturization of transistors and memory elements, and large bandwidth communication links powered by semiconductor lasers has also led to unexpected outcomes, such as an explosion of social media.

The first semiconductor "point-contact" transistor of 1947 was a few centimeters in size. Fig. 1.4 shows this device in comparison to a much newer transistor – the Silicon FinFET of 2012. The transistor is so small today that you can count the number of atoms across the fin – the scale has shrunk by a factor 10^7 from centimeter in 1947 to nanometer in 2012. In the latest generation microprocessors, several billion transistors can fit into a cm-size chip. The orders-of-magnitude increase in computational power and storage capability is a direct result of this miniaturization. Such orders of magnitude improvements over sustained periods of time is extremely rare in the recorded history of science.

Where are things going next? Some semiconductor experts and accomplished practitioners in this art may tell you that the golden era of semiconductors is over. That transistors have reached such small dimensions and semiconductor lasers are so efficient that there is not much more one can do. Do not buy into such predictions, because most such predictions can be proven to be wrong[3]. Come back and read this chapter once more. Just like it has happened again and again, some of you will do experiments or come up with a theory that will make a mockery of all current belief of what is possible with semiconductors and related materials. The future of our civilization depends on this adventure that you must embark upon! This book is an effort to motivate you to take this journey by arming you with the science that will serve as your fuel.

1.4 These boots are made for walking

This book is meant to be read with enjoyment and wonder, the concepts discussed in and outside of class, in office hours, and over coffee. Every effort will be made to present the development of the field as a uniquely human endeavor – laws and theorems did not fall out of the sky, but are creations of precise experimental observation and recognition of mathematical patterns – both achieved by human ingenuity. Please email me (`djena@cornell.edu`) your feedback and suggestions, as they go a long way in making sure that the lofty goals of this book are indeed achieved.

The book is divided into four modules, or parts:

- Part I: **Fundamentals:** In this module, in Chapters 1 through 7, we develop the basic concepts of quantum mechanics and quantum statistics, with notions of perturbation theory to apply to semiconductor physics.
- Part II: **Bands, Doping, and Heterostructures:** In Chapters 8 – 15, the core concepts of electronic states in semiconductors, heterostructures, and nanostructures are developed, culminating in

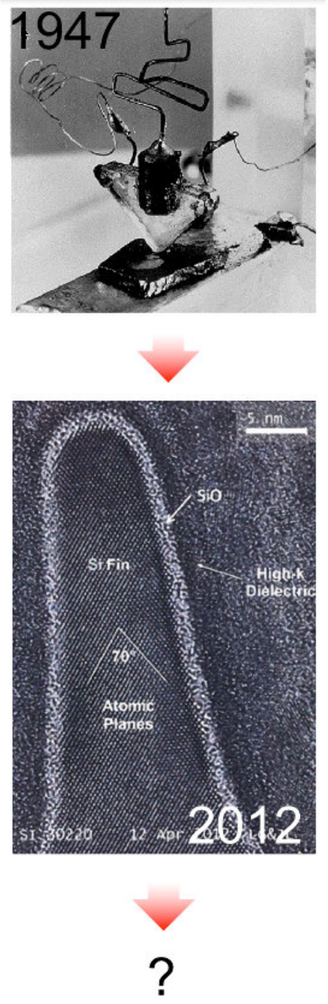


Fig. 1.4 The first centimeter-scale point-contact transistor, and today's nanometer-scale FinFETs.

---

[3] "It is exceedingly difficult to make predictions, especially about the future" – Danish proverb, sometimes attributed to Niels Bohr invoking the uncertainty principle in his answer to the question – "What will be the impact of quantum mechanics on the future of the world?"

Chapter 16 that summarizes, and connects the first half of the book to the next.

- Part III: **Quantum Electronics:** In Chapters 17– 25, the quantum concepts are applied to semiconductors and nanostructures to explain several electronic phenomena, with applications to diodes, memory, and transistors.
- Part IV: **Quantum Photonics:** In chapters 26– 29, light-matter interaction is discussed, and the quantum processes responsible for solar cells, light emitting diodes (LEDs), and lasers are discussed, and Chapter 30 discusses a possible future of the field.

1.5 Chapter summary section

At the end of each chapter, there will be a boxed summary emphasizing the major concepts that we have learned:

- This book is meant to explain the physics of semiconductors, and connect the physics to several applications such as diodes, transistors and lasers.
- The book is divided into four modules: (I) Fundamentals, (II) Semiconductor Physics, (III) Quantum Electronics with Semiconductors, and (IV) Quantum Photonics with Semiconductors.
- We discussed how to approach the contents in the book. For effective learning, there is no alternative to working through the concepts in the chapters, and solving the Exercises at the end of each chapter. I have purposely shared historical contexts and anecdotes in the text with the hope that familiarity with the stories and personalities behind the science and technology will lead to a deeper understanding and appreciation of the subject.

Further reading

At the end of each chapter, there is a short section on recommended readings that amplify the discussions in the chapter:

Thirty Years that Shook Physics (George Gamow) gives a delightful and readable account of how quantum mechanics was developed, complete with a stage play at the end. For a deep insight into the development of semiconductor physics starting from quantum mechanics as the technologically successful arm of condensed matter physics, and for the persons and organizations instrumental in its development, **Crystal Fire** (Riordian and Hoddeson) and **True Genius** (Daitch and Hoddeson) are recommended. **The Story of Semiconductors** (Orton) traces the development of semiconductors and their science and technology up to the present time. **The Physics of Semiconductors** (Grundmann) has an excellent timetable of semiconductors, and broadly is an excellent companion for the topics covered in this book. For those who want to understand how deep science and engineering, entrepreneurship, inspirational management, and a bit of luck goes hand in hand to commercialize technological breakthroughs in semiconductors, the following two books are highly recommended: **The Idea Factory** (Gertner), and **Fire in the Belly: Building a World-leading High-tech Company from Scratch in Tumultuous Times** (Neal and Bledsoe).

For details of each chapter-end reference, please refer to the end of the book.

Exercises

(1.1) **Cat's whiskers**
Given the preponderance of cat videos on the internet today, it is ironic that the same semiconductors that power today's web got their earliest break in industry in a device that used properties of the cat's whisker (Fig. 1.5). Research the web and write a short summary of how the earliest semiconductor diode rectifiers changed the radio, and the specific need for the physics of feline whiskers.

Fig. 1.5 Cat's whisker, the shape that launched the first semiconductor devices used for crystal-radios.

(1.2) **Moore's law**
In 1965, Gordon Moore working at Fairchild Semiconductor at the time (and to-be co-founder of Intel) made a prediction on the scaling of semiconductor transistor sizes and their packing density on chips with time. Though Moore's law is not a law of mathematics or physics such as the law of gravitation, it does capture an economic and computational argument that has remained approximately true for half a century. Perform some research and write a short summary of Moore's law, its implications, and its inevitable and ultimate demise.

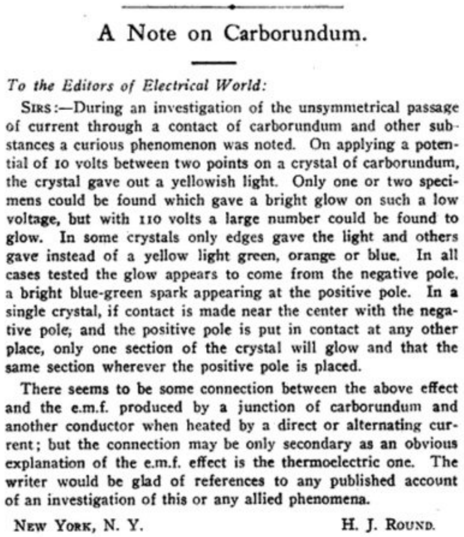
(1.3) **Mining semiconductor materials**
The most heavily used semiconductors today for high-performance electronic and photonic devices are silicon (Si), silicon carbide (SiC), germanium (Ge), gallium arsenide (GaAs), and gallium nitride (GaN). Write a short critique, discussing their abundance on earth and in the universe, and how some needs drive the search for alternative options.

(1.4) **Take two: John Bardeen**

Fig. 1.6 John Bardeen, a founding father of *both* semiconductor physics and of the physics of superconductors. He is the only person in history to have been awarded two Nobel Prizes in Physics: in 1956 for the invention of the semiconductor transistors, and in 1972 for the theory of superconductivity.

John Bardeen made pioneering contributions to the physics of semiconductors. He was a co-discoverer of the transistor with Brattain and Shockley. Perform research on his life and contributions to semiconductor physics, and write a paragraph summarizing them.

(1.5) **Let there be light**
In 1907, H. J. Round reported the observation of the emission of light from a metal/semiconductor rectifier. Read the (entire) report in Fig. 1.7, and do research on how the same principle of light emission from SiC remained the principal means of obtaining blue light at relatively low efficiencies from semiconductors till the breakthroughs in the 1990s in the gallium nitride semiconductor family revolutionized solid-state lighting.



A Note on Carborundum.

To the Editors of Electrical World:

SIRS:—During an investigation of the unsymmetrical passage of current through a contact of carborundum and other substances a curious phenomenon was noted. On applying a potential of 10 volts between two points on a crystal of carborundum, the crystal gave out a yellowish light. Only one or two specimens could be found which gave a bright glow on such a low voltage, but with 110 volts a large number could be found to glow. In some crystals only edges gave the light and others gave instead of a yellow light green, orange or blue. In all cases tested the glow appears to come from the negative pole, a bright blue-green spark appearing at the positive pole. In a single crystal, if contact is made near the center with the negative pole, and the positive pole is put in contact at any other place, only one section of the crystal will glow and that the same section wherever the positive pole is placed.

There seems to be some connection between the above effect and the e.m.f. produced by a junction of carborundum and another conductor when heated by a direct or alternating current; but the connection may be only secondary as an obvious explanation of the e.m.f. effect is the thermoelectric one. The writer would be glad of references to any published account of an investigation of this or any allied phenomena.

NEW YORK, N. Y. H. J. ROUND.

Fig. 1.7 Round's 1907 report on his surprising observation of the emission of light on applying voltage across a crystal of carborundum, which we now know is the compound semiconductor silicon carbide (SiC). This was the first semiconductor light-emitting diode or LED. LEDs made with the semiconductor gallium nitride (GaN) has revolutionized the lighting industry in the last decade.

(1.6) **Units and symbols: safety protocols!**
The limited number of alphabets in the English language makes it inevitable that some notations end up being used to represent more than one quantity. Also, different fields use different units of measure: and for good reason. Imagine a person working on semiconductors having to measure transistor gate lengths in light years, or a cosmologist having to state the distance between galaxies in nanometers! This problem alerts you to such instances in this book to avoid the pain or confusion they may cause when you encounter them.

(a) Symbols: The most painful choice I had to make in writing this book is for the symbol of the *magnitude* of the electron charge: e vs. q. The sign is another matter, which we discuss after the symbol. I chose q because the use of e as Euler's number $e = \sum_{n=0}^{\infty} \frac{1}{n!} = 2.718...$ in the the free electron wavefunction $\sim e^{ik \cdot r}$. The time of reckoning for this choice comes in later chapters (such as Chapter 23 on scattering and mobility), when the symbol $q = k_i - k_f$ is the difference of electron wavevectors before and after scattering. We take refuge by switching back to e for a few sections there. I hope this does not do too much violence. The reader can use her or his judgement with a lightning fast units check in case of confusion. Planck's constant is h, and "h-bar", or $\hbar = \frac{h}{2\pi}$. In photonics, frequency is represented by ν (in hertz), and "circular frequency" by $\omega = 2\pi\nu$ in cycles/s. This means $h\nu = \hbar\omega$ can be used interchangeably for energy. In electronics, f is preferred for frequency in hertz, and transistor speeds are stated as f_T and f_{max}.

(b) Signs: We owe this one to Benjamin Franklin for choosing the charge of the electron to be negative. In this book the charge of the electron is $-q$. Generally this should not be a problem, except when the minus signs appears in front of every expression, making the book look like a top secret military cryptogram. For almost every case, the signs in the text respect Franklin's choice. But there are at least one or two instances where I have committed the necessary sin of leaving the negative sign out when it does not hurt.

(c) Language: In this book, *bandstructure* and *band diagram* mean different things. The bandstructure is the $E(k)$ energy eigenvalue spectrum of a semiconductor. The energy band diagram is the spatial variation of the conduction band minimum $E_c(x)$ and valence band maximum $E_v(x)$. Under no instance should they be confused with one another.

(d) Units: We use SI units throughout. But in semiconductors, length is expressed in cm, microns, or nanometers depending on the natural scale of the physical quantity under consideration. For example, electron or hole densities are expressed in cm^{-3}, photon wavelengths in microns, and gate lengths of small transistors in nanometers.

(1.7) **Semiconductor cheat-sheets: where to find them**
(a) Tables A.1, A.2, and A.3 in the Appendix of this book listed at the end contain the physical properties of several semiconductors of the 3D and 2D form. Study these tables and use them.

(b) Table A.4 in the Appendix has the numerical values of physical constants that will be needed throughout the book.

(b) Fig. A.1 in the Appendix at the end of this book lists the constitutive equations that govern the quantum physics of semiconductors and their electronic and photonic device applications.

Secrets of the Classical Electron

2

2.1 Our ancestors knew metals 12
2.2 Discovery of the electron and its aftermath 12
2.3 Drude's model explains Ohm's law 12
2.4 Metals are shiny 15
2.5 Metals conduct heat 16
2.6 Icing on the cake: The Wiedemann–Franz law 17
2.7 All is *not* well 18
2.8 Chapter summary section 19
Further reading 20
Exercises 20

I would like to emphasize from the beginning how experiments have driven the search for new theories. New theories are born when experimental facts defy the sum total of all existing theoretical models. This happens again and again because when you perform an experiment, you cannot stop a physical law from doing its thing. Meaning, any experiment ever performed has measured every *known* law of physics, and every *yet to be discovered* law of physics! So when you measure something in regimes no one has ever done before, because of better equipment, or because you are just plain clever, new physical phenomena reveal themselves to you if you are a careful researcher. We will repeatedly observe these sorts of dynamics at play.

Because understanding comes from a quantitative theory that can explain experiments, and a quantitative theory is written in the language of mathematics, at times experimental observations languish till the development of the required new mathematics. For example, Newton had to *develop* the mathematical language of calculus to quantitatively explain the forces that drive the experimentally measured paths of planets. In a similar vein, for a quantum mechanical understanding of semiconductors and nanostructures, some notions of matrix based mathematics and Hilbert spaces will be needed in the book. But that can wait, as we can say so much by resting on the shoulders of Newton, Boltzmann, and Maxwell.

In this chapter we glimpse how the observed electronic, thermal, and photonic properties of metals were explained in the pre-quantum era using the three pillars (Fig. 2.1) on which physics rested around 1900s: Newton's **Classical Mechanics**, Boltzmann's **Thermodynamics**, and Maxwell's **Electromagnetism**. But then clever experimentalists pushed measurements to new regimes revealing physical phenomena that were in stark contradiction to these theories. Let's start talking about metals with the following questions:

- Why do metals conduct electricity and heat so well?
- Why are metals shiny?
- Why is the ratio of thermal and electrical conductivities a constant for most metals?

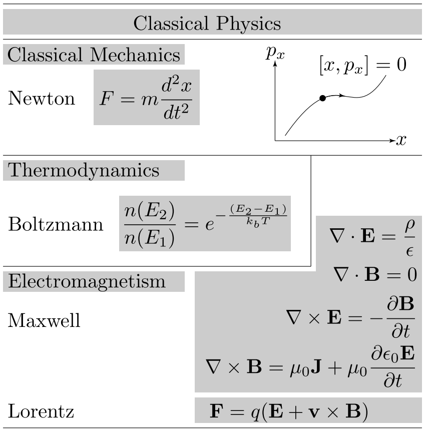

Classical Physics
Classical Mechanics
Newton $F = m\frac{d^2x}{dt^2}$
p_x
$[x, p_x] = 0$
x
Thermodynamics
Boltzmann $\frac{n(E_2)}{n(E_1)} = e^{-\frac{(E_2-E_1)}{k_bT}}$
Electromagnetism
Maxwell
$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon}$
$\nabla \cdot \mathbf{B} = 0$
$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \frac{\partial \epsilon_0 \mathbf{E}}{\partial t}$
Lorentz $\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$

Fig. 2.1 Three pillars of classical physics.

2.1 Our ancestors knew metals

Because nearly two-thirds of the elements in the periodic table are metals, and many occur in nature, they were discovered early, well before the 1900s. Metals have been known for a very long time[1] to be very different from insulators by being:

- good conductors of electricity,
- good conductors of heat, and
- reflective and shiny.

Wiedemann and Franz in 1850s experimentally discovered a deep and mysterious connection between the thermal conductivity κ, and the electrical conductivity σ of metals. They found that:

- The ratio $\frac{\kappa}{\sigma T} \sim 10^{-8}(\frac{\mathrm{V}}{\mathrm{K}})^2$, where T is the temperature, is a constant[2] for many different metals.

But why? Though there were several attempts to explain all these physical characteristics of metals, a truly satisfying explanation had to wait for the discovery of the single entity responsible for every one of the properties of metals discussed above: the **electron**[3].

2.2 Discovery of the electron and its aftermath

J. J. Thomson (Fig. 1.2) discovered the electron in 1897 in the Cavendish laboratory. He found the existence of a particle of mass $m_e = 9.1 \times 10^{-31}$ kg and electrical charge $-q$, where $q = 1.6 \times 10^{-19}$ Coulomb is the magnitude of the electron charge. This was followed by the discovery of the nucleus[4] by Ernest Rutherford in 1911, and the neutron by James Chadwick way later in 1932. In this chapter, the nucleus of the atom will play a cameo role of passive obstructions in the path of freely moving electrons in a metal. It will increasingly assert its role in later chapters. Today, the electron is known to be an elementary point-particle (of zero radius), and in addition to its mass and charge, also has an intrinsic angular momentum called the spin. Exactly how a point particle of no radius can have these properties is still under investigation, we will encounter their consequences in Chapter 3. In this chapter, we assess the consequences of its classical charge and mass.

Fig. 2.2 Paul Drude in 1900 proposed a model that combined classical mechanics, electromagnetism, and thermodynamics to explain the many properties of metals by invoking the then newly discovered electron.

2.3 Drude's model explains Ohm's law

The discovery of the electron was the trigger that precipitated an explanation of most properties of the metal discussed in Section 2.1. This was achieved by Paul Drude (Fig. 2.2) , who applied the notion that electrons move in a metal just like molecules move in a gas[5] follow-

---

[1] All these properties were known before the discovery of the electron, or the atom.

[2] This experimental fact is known as the empirical **Wiedemann-Franz law**.

[3] The word "electron" represents an indivisible portion of "electricity" – the word electricity predates the word electron by several centuries. It motivated a number of "..on" names: photon, proton, neutron, phonon, fermion, boson...

[4] Protons help maintain charge neutrality of the atom, and neutrons stabilize the nucleus and keep the atom from disintegrating. Protons and neutrons are *not* elementary particles like the electron; they are composed of quarks.

[5] Hence the name "electron gas".

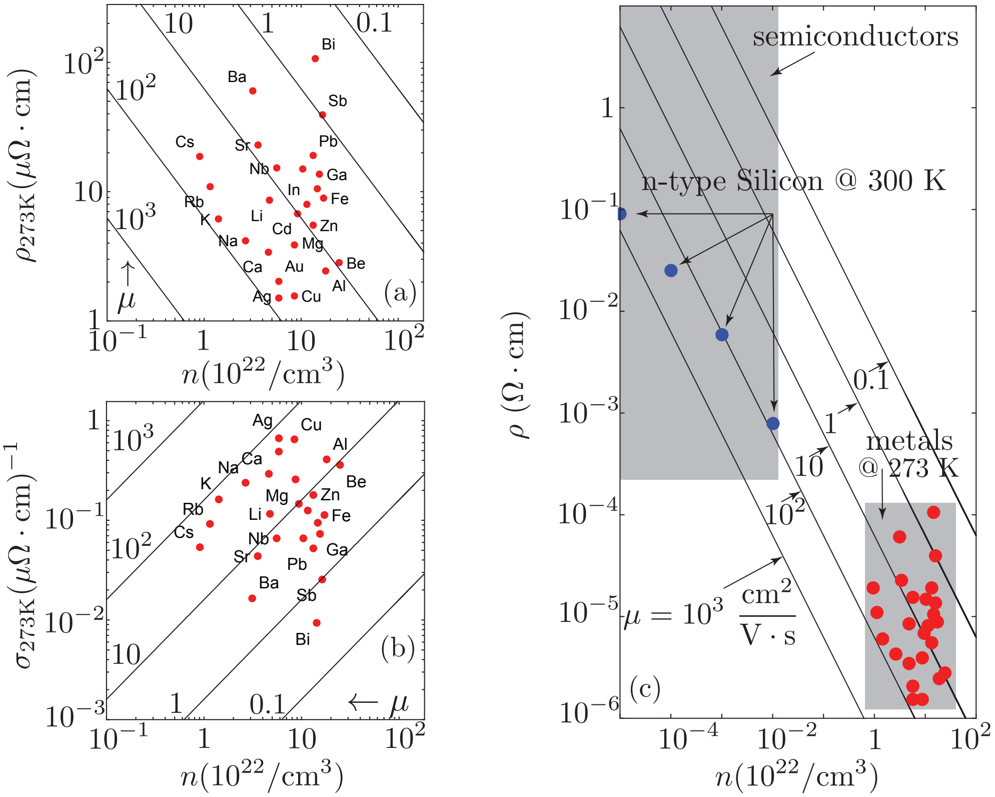


Fig. 2.3 (a) Electronic resistivity $\rho = 1/(qn\mu)$, and (b) conductivity $\sigma = 1/\rho = qn\mu$ of various metals at $T = 273$ K, and (c) a comparison of resistivities of metals to semiconductors as a function of mobile electron concentration n. The lines are of constant mobility μ in units of $\mathrm{cm}^2/(\mathrm{V{\cdot}s})$. Semiconductors typically have $\sim 2-10$ *orders* lower mobile carrier concentrations than metals. Unlike metals in which the mobile carrier concentration n is fixed, in a semiconductor n can be varied by various methods such as doping, field-effect, heat, or light. The variation in n shown here for Silicon is obtained by varying the concentration of Phosphorus doping. The data for metals is adapted from Solid State Physics by Ashcroft and Mermin.

ing the laws of classical mechanics, thermodynamics, and electromagnetism.

The first task was to explain the electrical conductivity σ of metals. Fig. 2.3 shows the measured resistivities and mobile carrier concentrations for several metals. The resistivities $\rho = 1/\sigma$ of most metals vary between $1 < \rho < 100$ $\mu\Omega$-cm. The experimental fact measured for metals is that the electric charge current I flowing through them is linearly proportional to the voltage V across them. This is stated as Ohm's law $V = IR$, where R is the *resistance* of the metal as shown in Fig. 2.4. Current is measured in amperes (or coulombs/second), voltage in volts, and resistance in ohms=volts/amperes. Drude imagined that the metal is filled with electrons of volume density in cm^{-3} units $n = \frac{N}{V}$, where N is the total number of electrons in the metal, and $V = AL$ is the volume as indicated in Fig. 2.5. The current density J measured in $\mathrm{amp/cm}^2$ is defined by $I = JA$. The expression for the

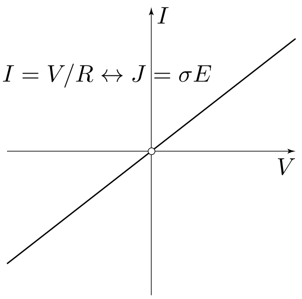


Fig. 2.4 Ohm's law is $V = IR$, or equivalently $J = \sigma E$.

current density is $\boxed{J = qnv}$.

$q \rightarrow$: The charge current density is given by the *particle flux density* nv times the charge of the particle, q. The flux density of particles is the volume density times the velocity, $n \times v$. Each electron in this flux[6] drags along with it a charge of q, so the current density is $J = qnv$. Later when we encounter heat, spin, or other currents, we will simply multiply the particle flux density with the corresponding quantity that is dragged along by the particle. In this book, we use the notation $q = +1.6 \times 10^{-19}$ coulomb. So the charge on the electron is $-q$, and on a proton is $+q$.

$n \rightarrow$: Because the structure of the atom was not known at the time, the electron density n is an empirical number, of the order $10^{23}/\text{cm}^3$ in Drude's model[7]. The density can, however, be experimentally measured by the **Hall effect**, which gives *both* the density n and the *sign* of the particles carrying current. We will see later that metals have *far more* electrons than $\sim 10^{23}/\text{cm}^3$ in them, with the number increasing as we go downwards in the periodic table to heavier elements. Most electrons however are stuck in core states or filled bands – concepts we develop in subsequent chapters as they require quantum mechanics to explain. Some electrons are free to wander around the crystal and conduct charge current – those are the *conduction electrons* that Drude's quantity n considers, the others simply do not play a role.

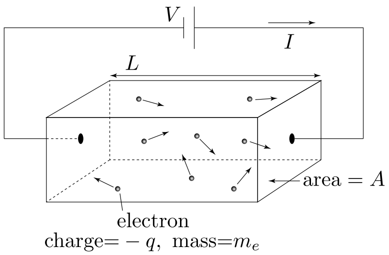


Fig. 2.5 Electron gas moving in response to an electric field in a metal.

$v \rightarrow$: Drude found the *average* velocity of the electrons in the metal with the following argument. The force on the electrons is due to the **electromagnetic**, or Lorentz force $\mathbf{F} = (-q)(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, where $\mathbf{E}$ is the electric field, and $\mathbf{B}$ is the magnetic field. If $\mathbf{B} = 0$, let's assume the scalar form for simplicity where $|\mathbf{E}| = E = -\nabla V = -dV/dx = V/L$ is the electric field exerted on the electrons by the battery. Because of this force, electrons accelerate according to Newton's law $F = \frac{dp}{dt}$, where $p = m_e v$ is the electron momentum. But as they speed up, sooner or later they will bump into the atoms in the metal, just as a swarm of bees drifting blindly through a forest. Whenever such a collision occurs, the momentum of the electron is randomized, and it starts accelerating along the direction of the force again. The modified Newton's law accounting for the *dissipation* or damping of momentum every τ seconds on an average is:

$$(-q)E = m_e \frac{dv}{dt} - \frac{m_e v}{\tau}, \tag{2.1}$$

where the increase in momentum due to the force is tempered by the decrease upon collisions every τ seconds[8]. If we ask what happens in the "steady state", meaning we have applied the voltage and waited long enough that all transients have died out, we can use $\frac{d}{dt}(...) \rightarrow 0$, which yields the velocity v as

$$v = \frac{q\tau}{m_e} E = \mu E \implies \boxed{\mu = \frac{q\tau}{m_e}.} \tag{2.2}$$

---

[6] Why does every electron in the universe have *exactly the same* charge? Paul Dirac, who we will encounter in the next chapter offered a delightful explanation for this observation based on quantum mechanics and topology, but it requires the existence of a magnetic monopole, which is yet to be observed!

[7] In this book, we will exclusively use SI units. In semiconductor physics, the length is often in cm instead of m. For example, particle densities are in $/\text{cm}^3$, and electric field in V/cm.

[8] The electrons impart momentum to the atoms upon collision. At high current densities this momentum transfer can physically move the atoms. This phenomenon is called *electromigration*, and is a real problem in microprocessors.

The electrons achieving an ensemble steady state velocity in the presence of a constant force is similar to a parachutist reaching a terminal velocity in spite of the constant gravitational force[9]. The drift velocity is proportional to the electric field, and the constant of proportionality $\mu = \frac{q\tau}{m_e}$ is defined as the **mobility** of electrons[10]. The concept of mobility will prove to be more useful for semiconductors than metals, as we will see later. It is clear that if the electron scatters less often, $\tau \uparrow \implies \mu \uparrow$.

Now putting together all the above pieces, Drude found that the current density is

$$J = qnv = \frac{nq^2\tau}{m_e}E = \sigma E \implies \boxed{\sigma = \frac{nq^2\tau}{m_e},} \tag{2.3}$$

where it is seen that the current density is proportional to the electric field, with the proportionality constant $\sigma = \frac{nq^2\tau}{m_e}$ called the electrical **conductivity**. The components appeal to our intuition - more electrons, and longer scattering times should lead to higher conductivity. Electrical conductivity does *not* depend on the *sign* of the charge q.

To close the story, we revert back to the current and see how Drude's model explained Ohm's law of the electrical conductivity of metals:

$$I = JA = \sigma EA = \sigma \frac{V}{L} A \implies V = I \cdot \frac{1}{\sigma}\frac{L}{A} = IR \implies \boxed{R = \frac{1}{\sigma}\frac{L}{A} = \rho\frac{L}{A}.} \tag{2.4}$$

The resistance R is measured in ohms. It is related to the microscopic details provided by Drude via the conductivity σ or the resistivity $\rho = \frac{1}{\sigma}$, and to the macroscopic dimensions[11] via L and A. A longer metal has more resistance because it takes an electron longer to get through due to more scattering obstructions. This is the essence of classical mechanics applied to the electron.

2.4 Metals are shiny

Why do metals reflect light? The secret lies in the swarm of conduction electrons in them! Maxwell (Fig. 2.6) had shown that a beam of light is a wave of oscillating electric and magnetic fields. The time dependent electric field of a light beam of amplitude E_0 is $E(t) = E_0 e^{i\omega t}$, where the circular frequency $\omega = ck = c\frac{2\pi}{\lambda} = 2\pi f$ is linked to the speed of light c and its color, or in other words its wavelength λ. If we subject the swarm of electrons in the metal not to the constant DC voltage of a battery as we did in Section 2.3, but to this new pulsating field, we can explain why metals reflect light. Instead of me solving this problem, can you do it for yourself? Sure you can. I have outlined the method in Exercise 2.5 at the end of this chapter. In a few steps you can explain the luster of metals, and also the colors of some!

Fig. 2.6 James Clerk Maxwell in 1865 unified electricity and magnetism. By introducing the concept of the displacement current, he showed that light is an electromagnetic wave. He made significant contributions to several other fields of physics and mathematics.

[9] A more apt analogy is the average velocity of a car in a stop-and-go traffic since electrons accelerate every time between collisions.

[10] Since the mobility is the ratio of the velocity to electric field $\mu = v/E$, its unit as used in this book is (cm/s) / (V/cm) = cm^2/V·s.

[11] This concept works well for large resistors. But we will see later that this picture will fail spectacularly when the L and A become very small, comparable to the wavelength of electrons. These anomalies were observed in 1960s leading to the observation of quantized conductance.

2.5 Metals conduct heat

Drude's model of electrical conductivity leads naturally to an explanation of the thermal conductivity of metals. As a rule of thumb for metals, good electrical conductors are also good thermal conductors. For example, copper has one of the highest electrical conductivities at $\sigma = 1/\rho \approx 1(\mu\Omega \cdot \text{ cm })^{-1}$, and a high thermal conductivity of $\kappa \approx 4$ W/cm·K. Atoms likely play a small role in the thermal conductivity of metals because the density of atoms are similar in metals and insulators. Because typical electrical insulators are also poor thermal conductors[12], the thermal conductivity of metals must be due to the conduction electrons.

Based on this reasoning, a model for the thermal conductivity of metals due to electrons alone goes like this: consider a piece of metal as shown in Fig. 2.7. The volume density of electrons $n = \frac{N}{V}$ is uniform across the metal, but the left end is held at a hotter temperature T_1 than the right end, which is at T_2. There is a temperature gradient across the metal from the left to the right. Drawing analogy to the charge current density $J_{charge} = \sigma \cdot (-\nabla V)$, where the potential gradient $-\nabla V = E$ is the electric field, and σ the electrical conductivity, we hunt for an expression for the heat current in the form $J_{heat} = \kappa \cdot (-\nabla T)$, where ∇T is the temperature gradient. If we can write the heat current in this form, we can directly read off the thermal conductivity κ. Because the heat current density J_{heat} has units $\frac{\text{J}}{\text{cm}^2\text{s}} = \frac{\text{Watt}}{\text{cm}^2}$, the thermal conductivity κ must have units of $\frac{\text{Watts}}{\text{cm}\cdot\text{K}}$.

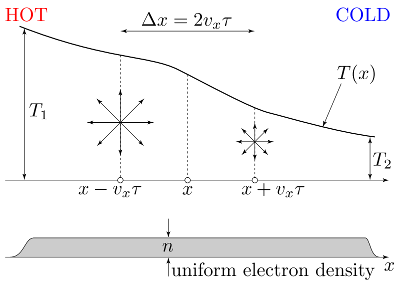


Fig. 2.7 Figure showing how a uniform density of electrons n in a metal can transport heat energy from the hot to the cold side.

Looking at Figure 2.7, we zoom into the plane x, and ask how much heat energy current is whizzing past that plane to the right. Because the density of conduction electrons is uniform in space $n(x) = n$, the heat current flows because of a difference in energy via the temperature gradient. Let $\mathcal{E}$ be the individual electron energy. The energy depends on x via the temperature of electrons at that plane. We shall call the average length traversed by the electron between scattering events the **mean free path**. In the plane one mean free path $v_x\tau$ to the left of x, the energy of electrons is $\mathcal{E}[T(x - v_x\tau)]$. Electrons in this (hotter) side have energies higher than those one mean-free path to the right of plane x, which have the energy $\mathcal{E}[T(x + v_x\tau)]$. Half the carriers at $x - v_x\tau$ are moving to the right, carrying a heat current density $\frac{n}{2}v_x\mathcal{E}[T(x - v_x\tau)]$. Similarly, half the carriers at $x + v_x\tau$ transport heat current $\frac{n}{2}v_x\mathcal{E}[T(x + v_x\tau)]$ to the left. The net heat current at x is the current flowing to the right, minus the current flowing to the left:

$$J_{heat} = \frac{n}{2}v_x[\mathcal{E}[T(x - v_x\tau)] - \mathcal{E}[T(x + v_x\tau)]]. \tag{2.5}$$

Now if we assume that the mean free paths $v_x\tau$ are small, we can write the heat current as

$$J_{heat} = \frac{n}{2}v_x\frac{\Delta\mathcal{E}}{\Delta T}\frac{\Delta T}{\Delta x}\Delta x = \frac{n}{2}v_x\frac{d\mathcal{E}}{dT}\frac{dT}{dx}(-2v_x\tau). \tag{2.6}$$

We use the fact that because of isotropic motion of electrons in three

[12] Diamond, BN, SiC, and AlN are exceptions, to be discussed later.

dimensions, $v_x^2 = v_y^2 = v_z^2 = \frac{v^2}{3}$. To obtain the term $\frac{d\mathcal{E}}{dT}$, we connect it to the definition of the electron specific heat, which is given by $c_v = \frac{1}{V}\frac{d\mathcal{U}}{dT}$, where $\mathcal{U} = N\mathcal{E}$ is the total energy of the electron system, and $n = \frac{N}{V}$. The physical meaning of the specific heat is the amount of energy that must be provided into the electron system to cause a unit degree rise in its temperature. For free classical motion of the electron gas in three dimensions, classical thermodynamics asserts an equipartition of energy, prescribing $\frac{1}{2}k_bT$ of energy for motion in each degree of freedom, or dimension of motion. Here $k_b = 1.38 \times 10^{-23}$ J/K is the Boltzmann constant, the fundamental quantity underpinning classical thermodynamics. So $\frac{1}{2}m_e v_x^2 = \frac{1}{2}k_bT$, and the total kinetic energy of each electron is then $\mathcal{U} = \frac{1}{2}m_e v^2 = \frac{1}{2}m_e(v_x^2 + v_y^2 + v_z^2) = \frac{3}{2}k_bT$. The specific heat of the electron gas is then seen to be $c_v = \frac{1}{V}\frac{d(N\times\frac{3}{2}k_bT)}{dT} = \frac{3}{2}nk_b$. We can write the heat current density as:

$$\boxed{J_{heat} = \frac{1}{3}c_v v^2 \tau(-\nabla T) = \kappa(-\nabla T).} \tag{2.7}$$

In the form of Equation 2.7, the Drude model of heat current carried by electrons in a metal gives us a thermal conductivity $\kappa = \frac{1}{3}c_v v^2\tau$ by analogy to the charge current. But the Drude model does more: it also gives a quantitative explanation for the ratio of electrical and thermal conductivity.

2.6 Icing on the cake: The Wiedemann–Franz law

One of the major successes of the Drude electron model of electrical and thermal conductivity of metals achieved was to provide an explanation for the Wiedemann–Franz law, which had languished for half a century without an explanation.

From Equations 2.3 and 2.7, we get the ratio

$$\frac{\kappa}{\sigma T} = \frac{(\frac{1}{3}c_v v^2\tau)}{(\frac{nq^2\tau}{m_e})T} = \frac{(\frac{1}{3}\frac{3}{2}nk_b\frac{3k_bT}{m_e}\tau)}{(\frac{nq^2\tau}{m_e})T} = \frac{3}{2}(\frac{k_b}{q})^2 \implies \boxed{\frac{\kappa}{\sigma T} = \frac{3}{2}(\frac{k_b}{q})^2 = \mathcal{L}.} \tag{2.8}$$

Rather amazingly, every microscopic detail specific to a particular metal such as the electron concentration n and the scattering time τ cancels out in the ratio above! What remains are the two fundamental constants that underpin classical physics: the Boltzmann constant k_b, and the electron charge q. This seemed to explain the Wiedemann–Franz law beautifully (see Fig. 2.8). The ratio is called the classical Lorenz number $\mathcal{L}$ with a value $\sim 10^{-8}(\frac{\text{V}}{\text{K}})^2$. Thus, Drude's model of a classical electron gas seems to have resolved the mystery we set out with in Section 2.1. Or has it?

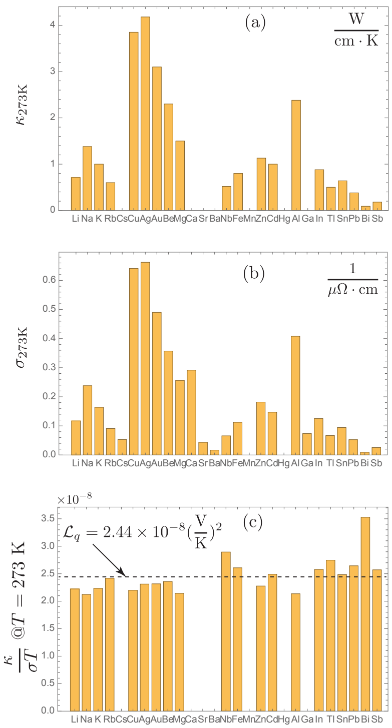


Fig. 2.8 (a) The thermal conductivity of various metals, (b) the corresponding electronic conductivity, and (c) the demonstration that $\kappa/(\sigma T)$ is the same for metals. The *quantum* Lorenz number $\mathcal{L}_q = \frac{\pi^2}{3}(\frac{k_b}{q})^2$ is shown as the dashed line, highlighting the Wiedemann-Franz law. The quantum Lorenz number is roughly twice of the classical value of Equation 2.8, and is discussed later.

2.7 All is *not* well

The success of the Drude model in explaining the Wiedemann– Franz law is remarkable, but unfortunately it is fundamentally flawed. Note that it does the best job with all the experimental data that was available till that time. We will see later that there is a crucial cancellation of two unphysical quantities that leaves the ratio intact. When experimentalists measured the specific heat c_v of electrons in metals, it was found to be *much smaller* than the classical result $c_v = \frac{3}{2} n k_b$ obtained by Drude. The experimental value is found to be roughly 1/60 of the classical result. This is a null result[13] that hints at something deeper – and we will see soon that something is **quantum mechanics**.

We will see in subsequent chapters that by demanding that electrons follow the rules of quantum mechanics instead of Newtonian mechanics, the electronic heat capacity was shown by Arnold Sommerfeld to be $c_v = [\frac{\pi^2}{2} n k_b] \cdot [\frac{k_b T}{E_F}]$ instead of the classical $c_v = \frac{3}{2} n k_b$. Here $E_F = \frac{p_F^2}{2m_e}$ is the **Fermi energy**, $p_F = \hbar k_F$ the **Fermi momentum**, and $k_F = (3\pi^2 n)^{\frac{1}{3}}$ is the **Fermi wavevector**. These three quantities are quantum-mechanical concepts, and simply cannot be explained by classical physics. Aside from the constants in the correct expression for the electron heat capacity, only a *small fraction* of the n conduction electrons: $\sim \frac{k_b T}{E_F}$ to be precise, seem to actually contribute to the heat capacity. We now know the reason: electrons being fermions, are subject to the **Pauli exclusion principle**. This principle prevents two electrons from occupying the same quantum state in the metal.

The consequence of the exclusion principle is rather drastic on the distribution of electrons. Fig.2.9 highlights this difference. In classical statistical physics, the number of electrons at an energy E goes as $e^{-E/k_b T}$, which is the Maxwell–Boltzmann distribution. So most electrons will pile up at the lowest allowed energies $E \to 0$ to lower the system energy. When extra energy is pumped into the electron system by whatever means, electrons at all energies in the distribution have equal opportunity to increase their energy. This is what led to the classical heat capacity $c_v = \frac{1}{V} \frac{d(N \times \frac{3}{2} k_b T)}{dT} = \frac{3}{2} n k_b$.

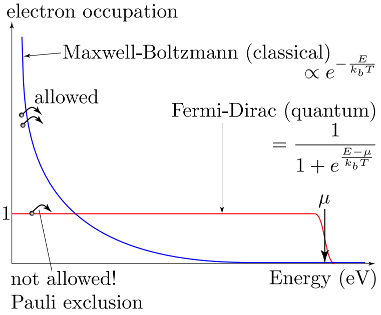


Fig. 2.9 Figure showing how the much higher heat capacity of electrons comes about because of assuming the classical Maxwell-Boltzmann distribution in energy. The Pauli exclusion principle of quantum mechanics changes the distribution to the Fermi-Dirac distribution, which fixes the electronic specific heat anomaly of the Drude model.

However, when electrons are forced to obey the Pauli exclusion principle, when an allowed state is occupied by an electron, it is impossible for a second electron to occupy the same state. The second electron must occupy a higher energy state. If we continue filling the states till the last electron, the highest energy occupied at $T = 0$ K is referred to as the **chemical potential** μ. While the chemical potential does not depend on temperature, the **Fermi energy** $E_F(T)$ does: it captures the increase in the energy of the electron distribution with heat. The two energies are equal at absolute zero, $\mu = E_F(T = 0)$. The occupation probability of a state of energy E upon enforcing the Pauli exclusion principle is the Fermi–Dirac distribution, $f(E) = 1/(1 + e^{\frac{E-\mu}{k_b T}})$, whose maximum value is 1. This is shown in Fig. 2.9. We will discuss this

[13] Huxley: The great tragedy of science: the slaying of a beautiful hypothesis by an ugly fact.

distribution in significant detail later.

The Fermi–Dirac distribution of electrons makes it clear why only a fraction $\frac{k_b T}{E_F}$ of electrons in a metal can actually absorb energy and promote themselves to higher energy states. Because of the Pauli exclusion principle, none of the electrons in the energy window $0 < E \lessapprox E_F - 3k_b T$ can increase their energy by absorbing $k_b T$ energy, because they are Pauli-blocked. Electrons in only a tiny sliver of energies $k_b T$ around the Fermi energy E_F have the freedom to jump to higher energy states that are unoccupied. Thus, the electronic heat capacity c_v is much smaller than what was predicted by the Drude model.

We will also see in Exercise 2.3 that the *sign* of the current carrying charge as measured by the **Hall effect** should always be negative for the charge of electrons, but it is measured to be positive for some materials. This flipping of sign also needs quantum mechanics to explain, as it is a result of the formation of **energy bands**. Finally, the *best* thermal conductor is diamond, with $\kappa \sim 22$ W/cm·K, roughly five times of copper, *and* it is an excellent electrical *insulator*! This goes smack at the face of the Wiedemann–Franz law because heat in some materials is not carried by electrons, but by lattice vibration quanta called **phonons**. In the next chapter, we begin our quantum journey and see why electrons must follow the exclusion principle and the Fermi–Dirac distribution in the first place.

2.8 Chapter summary section

In this Chapter, we learned:

- The electronic and thermal conductivities of metals were explained by the Drude model by attributing these properties correctly to the newly discovered particle, the electron.
- In the Drude model, free electrons subject to the laws of classical mechanics, electromagnetism, and thermodynamics could explain the electronic conductivity σ, and the thermal conductivity κ reasonably successfully.
- The Drude model also seemed to resolve a long-standing mystery of the empirical Wiedemann–Franz law, which stated the ratio $\frac{\kappa}{\sigma T}$ is a constant for metals.
- The heat capacity of free conduction electrons in metals predicted by the Drude model turned out to be inconsistent with the measured values, which were several orders too small. It would need the full machinery of quantum mechanics and a quarter century to resolve this discrepancy.

Further reading

For insightful and comprehensive discussion of the theory of metals, the classic text **Solid State Physics** by Ashcroft & Mermin remains unmatched in covering everything that was known till the early 1980s, and is highly recommended for critical reading. **The Theory of the Properties of Metals and Alloys** by Mott & Jones, a much older text, shows the earliest application of quantum mechanics to the understanding of metals and alloys. For a historical review of how the understanding of metals developed in its early times, **Reviews of Modern Physics, vol. 59 pg. 287 (1987)** by Hoddeson, Baym, and Eckert is an enjoyable read. Finally, the classic article that introduced the electron theory of metals to the world in its complete form, **Electronentheorie der Metalle** by Sommerfeld and Bethe, in Handbuch der Physik 1933 remains one of the most comprehensive to date. It is in German, and still worth a scan even if you don't know the language!

Exercises

(2.1) **Metals, insulators, and stars in the universe**

(a) Recent estimates put the number of stars in the universe at $\sim 10^{30}$, dark matter notwithstanding. Show that the span of electrical resistivities $\rho = \frac{1}{\sigma}$ in the best metals to the best electrical insulators is larger than the number of stars in the universe. Electrical resistivity of the best metals are in the range of $\rho \sim 10^{-6}$ ohm·cm.

(b) It is useful to find combinations of fundamental constants that yield physical quantities. Show that for electrical resistivity, such a ratio is $\rho_Q = \frac{h}{q^2} a_B^0$ where h is the Planck's constant, q the electron charge, and a_B^0 the Bohr radius of the Hydrogen atom. Show that the value of this "quantum" resistivity is roughly 100 times higher than the lowest resistivity metals.

(2.2) **Superconductivity**

In 1911, Heike Kamerlingh Onnes discovered that some metals when cooled to low temperatures lose their resistivity completely, meaning $\rho = 0$ exactly. The near-magical properties of superconductors are the source of as much endless fascination for scientists as for the general public. For example, they expel all magnetic fields in them, and float in thin air defying gravity. The levitation images (Fig. 2.10) are near iconic and trains have already been built on this concept. In this book, we will not discuss the physics of superconductivity, which has to do with the pairing of electrons to form a highly correlated state, which makes the current flow without dissipation. Unfortunately to date superconductivity has been observed only at low temperatures. But we sure can dream! Imagine tomorrow morning you wake up to the news that a room-temperature superconductor has been discovered: that news and its implications will surely change the face of the earth. What would you do with it?

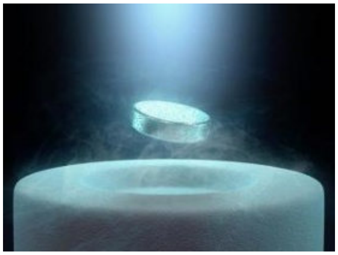


Fig. 2.10 Levitating superconductor.

(2.3) **Who's in charge? The Hall effect**

In 1879, nearly twenty years before the discovery of the electron, Edwin Hall (Fig. 2.11) during his

PhD work at the Johns Hopkins University made a most useful discovery. In Chapter 25, a detailed treatment of the Hall effect and behavior of electrons in magnetic fields is given. This Exercise is an early introduction to these phenomena.

Fig. 2.11 Edwin Hall discovered what is now known as the Hall effect in 1897 by detecting a lateral voltage in the current carrying gold leaf due to a magnetic field. This effect is now a heavily used method to measure the carrier concentration and mobility of current-carrying particles in all sorts of materials. Several new Hall effect variants have been uncovered, such as the Quantum Hall effect, the spin-Hall effect, etc. These are described in Chapter 25.

Consider a conductor of thickness L_z, length L_x, and width L_y respectively. A current I flows along the length, and a magnetic field B_z is applied along the thickness. The Lorentz force due to the electric and the magnetic field acts on the charges carrying the current.

(a) Show that if the volume density of charge carriers in the conductor is n, then a voltage develops in the direction perpendicular to the current, and to the magnetic field. This voltage is called the Hall voltage

$$V_H = -\frac{I_x B_z}{qnL_z}. \tag{2.9}$$

(b) Show why the **sign of the Hall voltage** is an experimental measurement of the sign of the charge of the current carrying particles.
(c) Show that the ratio of the lateral electric field E_y and the magnetic field and current density is given by

$$\boxed{R_H = \frac{E_y}{j_x B_z} = \frac{1}{qn}}, \tag{2.10}$$

which is independent of the sample dimensions and the magnetic field. The Hall coefficient thus is a direct measurement of the mobile carrier concentration in the material. Also, this simple model suggests that the Hall coefficient does *not* depend on any details of scattering events, or the magnetic field. Though it is mostly correct, metals can demonstrate varying degrees of magnetoresistance, but do not deviate too far from Equation 2.10, making it a trustworthy measurement of the carrier density and the sign of its charge.

We will come back to the discussion of the Hall effect, and why even in materials in which electrons carry the current, we can experimentally measure a positive sign of the carriers!

(2.4) **Breakdown: Critical electric fields in insulators**
(a) Consider the parallel plate capacitor: two flat metal plates separated by an electrical insulator barrier of thickness t_b and dielectric constant ϵ_b. If a voltage V is applied across the metal plates, show that the electron sheet density σ_m at the metal/insulator interfaces is $q\sigma_m = \epsilon_b \frac{V}{t_b}$. This result, which is a statement of Gauss's law, can also be viewed as the electric field $F = V/t_b$ across a planar insulator layer is equal to the sheet charge density divided by the dielectric constant of the insulator $\boxed{F = \frac{q\sigma_m}{\epsilon_b}.}$ This simple result is of very high importance in field-effect transistors or FETs.
(b) If the critical electric field at which the insulator breaks down is F_{cr}, the maximum sheet density that the insulator can sustain around it is $\sigma_{cr} = \epsilon_b F_{cr}/q$. The best electrical insulators break down at fields $F_{cr} \sim 10$ MV/cm. Estimate the maximum charge such an insulator can control by field effect. This is a fundamental limit that shows why the carrier density in typical metals cannot be modulated by field-effect using an insulator.
(c) The book **Plastic Fantastic** by Reich describes the unearthing of a major scientific fraud where a researcher claimed to have created a metal out of an organic semiconductor by field effect, and this organic metal became a superconductor at low temperatures. The researcher had not paid attention to the currently known limits of part (b) of this Exercise. Read up about this affair and be aware of it!

(2.5) **Mirror mirror on the wall**
Believe it or not, coating glass with metal to make a mirror was a *big* technological breakthrough back in the day (quite possibly marking the birth of fash-

ion?). In this Exercise, we answer why metals are shiny – why they reflect most of the visible light incident on them. Not surprisingly, this has to do with the conduction electrons in the metal. Here is our mode of attack: when the light wave experiences a mismatch in the refractive index, part of it is transmitted, and part reflected. So we will ask Maxwell's equations to give us the reflection coefficient when a light beam is incident from air on a metal. If we find that this reflection coefficient is very high for visible wavelengths, we have succeeded in explaining why metals are shiny.

The reflection coefficient Γ_r will depend on the refractive index $\sqrt{\epsilon(\omega)}$ of the metal, which in turn will depend on how the conduction electrons respond to the oscillating electric field of the light beam. This is where Drude's free electron model of the metal – the same model that explained electrical and thermal conductivity – will help us through.

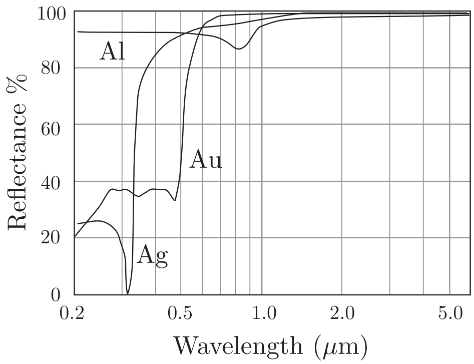


Fig. 2.12 Reflectance spectra of three metals (adapted from Wikipedia).

Fig. 2.12 shows the measured reflectance of three common metals as a function of the wavelength of light incident on it. Your job in this Exercise is to calculate and make your own plot by standing on Maxwell, Drude, and Newton's shoulders. If things go right, you may even explain the dips and wiggles in Fig. 2.12. I will roughly outline the method and trust you can finish the story:

The electric field of the light beam according to Maxwell oscillates in time as $E(t) = E_0 e^{i\omega t}$, where E_0 is the amplitude, and $\omega = 2\pi f$ is the radial frequency, $f = c/\lambda$ with c the speed of light, and the wavelength λ of light is the x–axis in the plot. The reflection coefficient for light is $\boxed{\Gamma_r = \frac{E_r}{E_i} = \frac{\sqrt{\epsilon_0} - \sqrt{\epsilon(\omega)}}{\sqrt{\epsilon_0} + \sqrt{\epsilon(\omega)}}}$, where $\sqrt{\epsilon}$ is the refractive index of the media. The reflectance is $R = |\Gamma_r|^2$, which can be found for various wavelengths; this is the y–axis of the plot. Note that all symbols have their usual meanings.

(a) From Maxwell's equation $\nabla \times \mathbf{H} = \mathbf{J} + i\omega\epsilon(\omega)\mathbf{E}$ in material media, show that the dielectric constant of the metal is $\boxed{\epsilon(\omega) = \epsilon_0[1 - i\frac{\sigma(\omega)}{\omega\epsilon_0}].}$

(b) Now if you have the frequency-dependent conductivity $\sigma(\omega)$, you can make your plot by looking up the properties of the metal! The DC Drude model for conductivity was obtained in this chapter as $\sigma(0) = \frac{nq^2\tau}{m_e}$. Here you need to use Newton's laws again for the AC electric field in the form $-qE_0e^{i\omega t} = m_e\frac{dv}{dt} - \frac{m_e v}{\tau}$ and solve to show the following:

$$\boxed{\sigma(\omega) = \frac{\sigma_0}{1 - i\omega\tau} = \underbrace{\frac{\sigma_0}{1 + (\omega\tau)^2}}_{\mathrm{Re}(\sigma(\omega))} + i\underbrace{\frac{\omega\tau\sigma_0}{1 + (\omega\tau)^2}}_{\mathrm{Im}(\sigma(\omega))}.} \tag{2.11}$$

(c) Now you are close to the finish line. Use the above modified Drude AC conductivity, look up the required properties of the three metals, and plot the reflectances of all the three metals. Compare with Fig. 2.12 and comment on the similarities and differences.

Quantum Mechanics in a Nutshell

3

3.1 Planck's photon energy quanta 23
3.2 Bohr's electron energy quanta 26
3.3 Wave-particle duality 28
3.4 The wavefunction 29
3.5 Operators 31
3.6 States of definite momentum and location 33
3.7 States of definite energy: The Schrödinger equation 35
3.8 Time-dependent Schrödinger equation 36
3.9 Stationary states and time evolution 37
3.10 Quantum current 39
3.11 Fermions and bosons 41
3.12 Fermion and boson statistics 44
3.13 The Spin-statistics theorem 46
3.14 The Dirac equation and the birth of particles 47
3.15 Chapter summary section 48
Further reading 49
Exercises 50

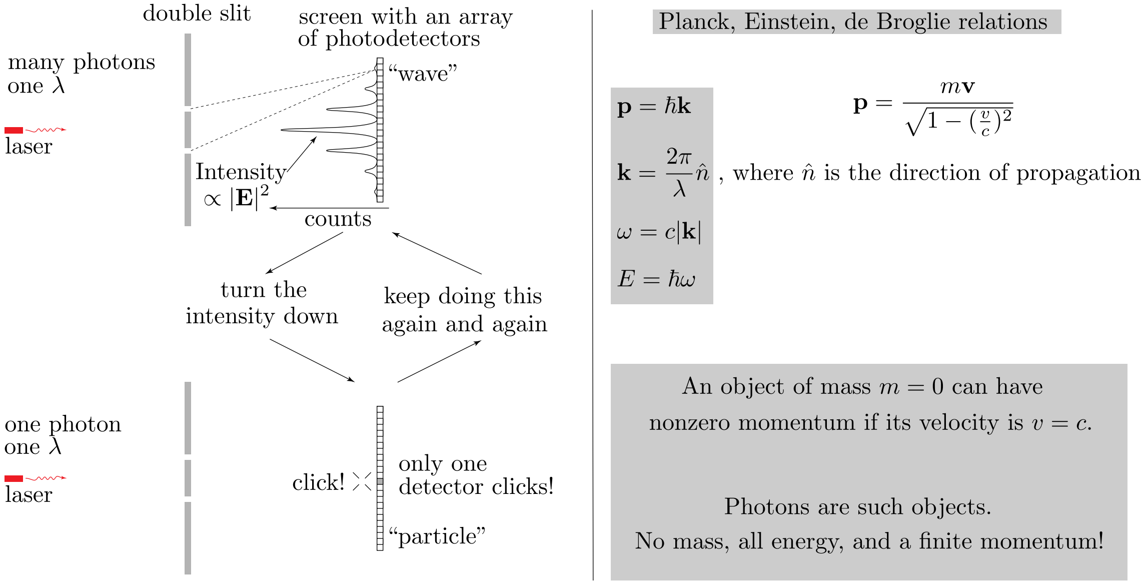
This chapter presents a very short summary of the major ideas of quantum mechanics. By tracing the historical development of its core ideas, we learn how to treat the static and the dynamic laws of electrons and their statistical properties by the laws of quantum mechanics rather than by classical thermodynamics and Newton's laws. In writing this chapter I thought I was kidding myself in literally trying to bottle a genie; the consequence is that this chapter is longer than the others. Quantum mechanics is one of the greatest intellectual advances in the history of science, and a single chapter can hardly do justice to the subject! I highly recommend that you append your reading of this chapter with your favorite quantum mechanics texts. I have suggested a few in the Further Reading section at the end of the chapter. The major goals of this chapter are to understand the following concepts:

- What experimental facts led to the development of quantum mechanics?
- What are the core ideas of quantum mechanics, and how are they different from classical mechanics?
- How did quantum mechanics lead to the explanation of the existence of atoms, molecules, and the entire periodic table?
- How did quantum mechanics forever change the core concepts of classical thermodynamics through the Fermi–Dirac and Bose–Einstein statistics?

3.1 Planck's photon energy quanta

Time: end of the 19th century. Maxwell's equations have established Faraday's hunch that **light** is an electromagnetic *wave*. However, experimental evidence mounted pointing towards the fact that a beam of light is composed of *particles* that pack a definite momentum and energy. Here is the crux of the problem: consider the double-slit experiment shown in Fig. 3.3. Monochromatic light of wavelength λ passing through two slits separated by a distance $d \sim \lambda$ forms a diffraction pattern on a photographic plate. If one tunes down the intensity of light in a double-slit experiment to a very low value, one does not get a "dimmer" interference pattern, but discrete strikes on the pho-

Fig. 3.1 Michael Faraday, one of the greatest experimentalists of all times. He discovered the relation between electric and magnetic fields, and influenced Maxwell to discover that light is an electromagnetic wave. Light played *the* central role in the development of quantum mechanics.


double slit
screen with an array of photodetectors
many photons
one λ
laser
"wave"
Intensity $\propto |\mathbf{E}|^2$
counts
turn the intensity down
keep doing this again and again
one photon
one λ
laser
click!
only one detector clicks!
"particle"

Planck, Einstein, de Broglie relations

$$\mathbf{p} = \hbar\mathbf{k} \qquad \mathbf{p} = \frac{m\mathbf{v}}{\sqrt{1-(\frac{v}{c})^2}}$$

$$\mathbf{k} = \frac{2\pi}{\lambda}\hat{n}\ , \text{ where } \hat{n} \text{ is the direction of propagation}$$

$$\omega = c|\mathbf{k}|$$

$$E = \hbar\omega$$

An object of mass $m = 0$ can have nonzero momentum if its velocity is $v = c$.

Photons are such objects.
No mass, all energy, and a finite momentum!

Fig. 3.3 Photons behaving as particles.

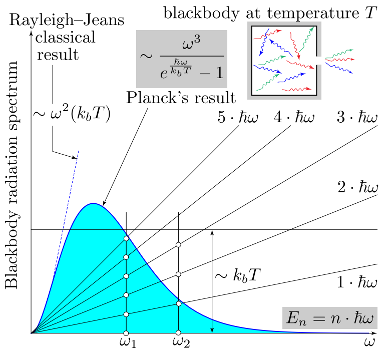


Fig. 3.2 The failure of classical physics to explain the experimental blackbody radiation spectrum in the high-energy regions (referred to as the ultraviolet catastrophe), and Planck's spectacular resolution of the catastrophe using the quantum hypothesis that the amount of energy of light in any mode ω is quantized and can exist only in packets of $E = n\hbar\omega$ where n is an integer.

tographic plate and illumination at specific points. This means light is composed of "particles" whose energy and momentum are concentrated in one point which leads to discrete hits. So here is **surprise 1**: Though light is a wave, its constituent packet is swallowed whole by just one detector! **Surprise 2** is the following: If the above experiment is repeated billions of times, and the count of each individual detector added up to create a histogram, the histogram becomes the classical diffraction pattern! This curious phenomenon means that though the energy and momentum of a packet of light is concentrated at a point and behaves as a particle, its wavelength extends over all space, which leads to diffraction patterns in the large number limit.

This experiment, routinely done today, was not possible in the end of the 19th century because experimental sources that could produce single light quanta (or photons) did not exist at that time. However, their existence was forced upon us as the only possible explanation of the **blackbody radiation** pattern: a well-known experimental fact for which classical physics could offer no explanation. If we measure the intensity of light radiated from a blackbody at a temperature T (for example, a cast-iron oven with a small opening in it), the radiation spectrum has a characteristic bell-shape shown as the experimental curve in Fig. 3.2. The spectral density $u(\omega)$ (or energy content) *decreases* at high frequencies ω (if it did not, we would not exist!). Rayleigh and Jeans had attempted to explain the spectrum. In Exercise 3.2, we will find that the number of modes (or the **density of states**) of light waves that fit in a box increases with frequency as ω^2. To each of these allowed modes of light, classical thermodynamics allocates $\sim k_bT$ of energy. This principle, derived in Chapter 4 is called the **equipartition of energy**. These considerations led Rayleigh and Jeans to the result $u(\omega) \sim \omega^2 (k_b T)$ for the energy content with frequency at a temperature T. Now this spectra diverges as ω^2 and blows up at high frequencies – the **ultraviolet catastrophe**. Clearly inconsistent with experiments, it remained one of the major mysteries of the time.

Max Planck (Fig. 3.4), after considerable internal struggle, postulated that in any allowed frequency mode ω, light energy can only exist in **quanta** of $\boxed{E = n(\hbar\omega)}$, where $n = 0, 1, 2, ...$ is an integer, and cannot assume a continuous range of energies. Here $\hbar = \frac{h}{2\pi} = \frac{6.63 \times 10^{-34}}{2\pi}$ J·s. h is the **Planck's constant**, and $\hbar$ is the **reduced Planck's constant**. In Fig. 3.2, Planck's quantum hypothesis is indicated schematically as the lines $E = n\hbar\omega$ crossing the origin. It is clear from this hypothesis that if energy $E \leq k_b T$ is available from heat as indicated by the horizontal line, the number of light quanta in any mode ω decreases with increasing $\hbar\omega$. For example, in the two modes shown, ω_1 has $n = 0, 1, 2, 3, 4, 5$ allowed, and ω_2 has $n = 0, 1, 2, 3$. It is clear that higher energy modes have less number of quanta occupied at a given energy because of the discrete number allowed for each mode.

This postulate led Planck to derive a spectral density of radiation given by $\boxed{u(\omega) \sim \omega^3 / (e^{\frac{\hbar\omega}{k_b T}} - 1}$). His formula **exactly** explained experimental data, resolving the ultraviolet catastrophe. In Exercise 3.2, I have pointed out steps for you to work this out for yourself, reliving Planck's struggle and the birth of quantum mechanics. Planck's key postulates were that photons in the mode of circular frequency ω can only have discrete energies $E = n\hbar\omega$, and each photon in that mode has a momemtum $\mathbf{p} = \hbar\mathbf{k}$. Here $\mathbf{k} = (2\pi/\lambda)\hat{\mathbf{n}}$, $\hat{\mathbf{n}}$ is a unit vector in the direction of propagation, and $\omega = c|\mathbf{k}|$ where c the speed of light. Planck's blackbody radiation formula has been experimentally confirmed with spectacular accuracy till today. It enables measurement of temperatures from microscopic to cosmic scales (see Exercise 3.11).

Just about five years after Planck's result, Einstein (Fig. 3.5) used it to explain the phenomena of ejection of electrons from metals by light known as the **photoelectric effect**, which is discussed further in Exercise 3.3. To explain the photoelectric effect, Einstein used two core concepts: (a) that the energy in any mode of light is quantized to $n(\hbar\omega)$ as postulated by Planck, and (b) invoked the recently discovered particle, the electron. Einstein was developing the theory of relativity around the same time. In this theory, the momentum of a particle of mass m and velocity v is $p = mv/\sqrt{1 - (v/c)^2}$, where c is the speed of light. Thus if a particle has $m = 0$, the only way it can pack a momentum is if its velocity is $v = c$. Nature takes advantage of this possibility and gives us such particles. They are now called photons. Thus photons have no mass, but have momentum. Thus light, which was thought to be a wave acquired a certain degree of particle attributes. So what about particles with mass – do they have wave nature too? Nature is too beautiful to ignore this symmetry!

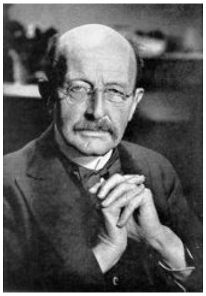


Fig. 3.4 Max Planck is considered to be the "father" of quantum mechanics. Postulated quanta of light (photons) to explain the blackbody radiation spectrum. He was forced to put forward this hypothesis because he discovered that it explained experimental data, in spite of the conflict it caused to him as he was a classically trained physicist where such quanta was inconcievable. Planck was awarded the Nobel prize in physics in 1918.

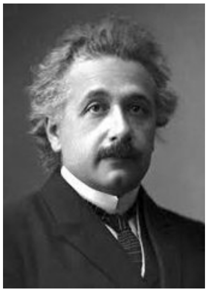


Fig. 3.5 Albert Einstein is considered the greatest physicist since Newton. In addition to special and general relativity, he contributed significantly to the development of quantum mechanics, and all areas of physics. He was awarded the Nobel Prize in 1921 for the photoelectric effect, an early experiment confirming quantum theory. He did not have a brother, and certainly did *not* start a bagel company.

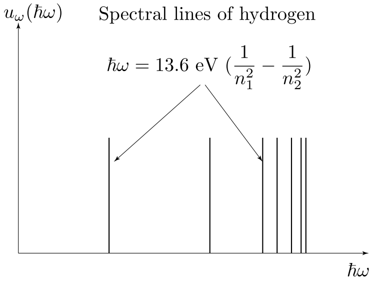


Fig. 3.6 The spectrum of pure Hydrogen is a set of sharp lines, in contrast to the blackbody radiation. The lines were precisely at energies given by the empirical Rydberg formula for which there was no explanation.

The interaction of light with matter (i.e. particles with mass) of its purest form – that is – with pure gases of the elements such as hydrogen resulted in spectra that was nothing like the blackbody spectrum seen in Fig. 3.2. Instead, the spectrum that was emitted by pure hydrogen was at a set of very peculiar and **sharp frequencies**. These frequencies were identified in 1880s by Rydberg to follow the empirical relationship

$$\hbar\omega = \underbrace{(13.6 \text{ eV})}_{Ry}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right), \tag{3.1}$$

where n_1 and n_2 are integers, as indicated in Fig. 3.6. The spectral lines were identified in sets with $n_1 = 1$ and $n_2 = 2, 3, ...$ in one set, $n_1 = 2$, and $n_2 = 3, 4, 5, ...$ in another, and so on. Though this strange empirical pattern was measured for several pure atomic species and the energy unit $Ry = 13.6$ eV came to be known as the Rydberg, the reason behind these sharp spectral lines remained a complete mystery. The solution to this mystery came from an unexpected corner; it had to wait for nearly three decades for the discovery of the heart of the atom itself.

3.2 Bohr's electron energy quanta

Fig. 3.7 Ernest Rutherford, the discoverer of the nucleus and the proton, is considered the father of nuclear physics. He was awarded the Nobel Prize in Chemistry in 1908.

In 1911, Ernest Rutherford (Fig. 3.7) discovered the nucleus of the atom: a very dense and small region in an almost empty atom that consisted of exactly the same net charge, but of sign opposite to that of the electrons. This discovery opened a literal pandora's box of questions regarding the stability of the atom. The simplest atom – hydrogen – has one electron and one proton in the nucleus. If these particles followed the laws of classical mechanics, the electron should instantly crash into the proton due to the immense Coulomb attraction force between them. In classical mechanics and classical electromagnetism, Earnshaw's theorem asserts that a collection of point charges cannot be in stable equilibrium due to electrostatic forces alone. So according to classical mechanics and electromagnetism, the hydrogen atom, or any other atom is unstable, and should not exist!

Fig. 3.8 Niels Bohr was the original architect and "conscience keeper" of quantum mechanics, and an intellectual leader who influenced an entire generation. He was awarded the Nobel Prize in 1922. His own theory certainly passed his own famous test: "We agree your theory is crazy. But it is not crazy enough to be true."

But we know from experiments that the stability of atoms is their hallmark: they are eternally stable objects. Immense energies of the scale of $E = mc^2$ are needed to break them apart. This conundrum was resolved by Niels Bohr (Fig. 3.8) with an explanation that shook classical mechanics to its core. It is important to go through his arguments to see how he resolves the conundrum, and for the first time is able to explain the origin of the sharp Rydberg spectral lines.

His arguments go as follows: the light beam incident on the atoms primarily interacts with the electrons as they are much lighter than the nucleus. The total energy E of the electron of mass m_e and non-relativistic momentum $p = m_e v$ a distance R from the proton in a hydrogen atom is the sum of the kinetic energy and the potential energy due to attraction to the nucleus:

$$E = \frac{p^2}{2m_e} - \frac{q^2}{4\pi\epsilon_0 R}. \tag{3.2}$$

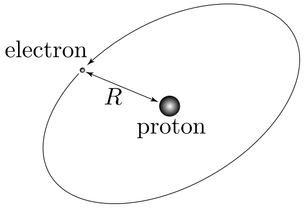
Now the classical values of the electron energy are *continuous* because R and p can assume any values. So the electron making a transition from a high energy state to a low energy state should emit a continuous spectra of photons, not the sharp discrete lines of Fig. 3.6. Because the experimental spectrum disagrees with classical energies, the *classical concept of all possible energies of Equation 3.2 must be incorrect*.

After several trials and errors, and specifically going back to the existing evidence of the spectral lines, Bohr found a radical solution. His singular hypothesis that solved the entire problem is this: the momentum of the electron p, and the path (or orbit) it takes around the proton in the hydrogen atom cannot assume the continuous values that are allowed by classical mechanics, but can only take values such that they satisfy the condition

$$\oint p dx = nh, \tag{3.3}$$

where h is Planck's constant, and $n = 1, 2, 3, ...$ is an integer, and $\oint ...$ is a closed loop integral over the circumference. If this is true, and we assume a circular orbit of radius R (Fig. 3.9), we must have

$$\oint p dx = p_n 2\pi R_n = nh \implies p_n R_n = n\hbar. \tag{3.4}$$



Hydrogen atom

Fig. 3.9 A classical picture of the Hydrogen atom.

Balancing the classical attractive Coulomb force with the classical centrifugal force we obtain that the electron can only assume discrete momentum values:

$$\frac{m_e v_n^2}{R_n} = \frac{q^2}{4\pi\epsilon_0 R_n^2} \implies p_n = \frac{q^2 m_e}{4\pi\epsilon_0\hbar} \cdot \frac{1}{n} = \frac{\hbar}{a_B} \cdot \frac{1}{n}, \tag{3.5}$$

and only orbits of fixed radius given by

$$R_n = n^2 \underbrace{(\frac{4\pi\epsilon_0\hbar^2}{q^2 m_e})}_{a_B} = n^2 a_B, \tag{3.6}$$

where $a_B = \frac{4\pi\epsilon_0\hbar^2}{q^2 m_e}$ is called the Bohr radius, and has a value of roughly 0.5 Å. Note that the picture is a set of allowed orbits of radii R_n in which the electron can move around the nucleus, just as the planets around the Sun in the Solar System[1].

Now going back to Equation 3.2 with these quantized momenta p_n and radii R_n, Bohr obtained the discrete energies

$$E_n = \frac{p_n^2}{2m_e} - \frac{q^2}{4\pi\epsilon_0 R_n} = \frac{1}{2}\frac{\hbar^2}{m_e a_B^2}\frac{1}{n^2} - \frac{\hbar^2}{m_e a_B^2}\frac{1}{n^2} = -\frac{\hbar^2}{2m_e a_B^2}\frac{1}{n^2} \tag{3.7}$$

[1] The theory is also called the "Bohr-Orbit" theory as a throwback to Newton's explanation of planetary orbits.

allowed for the electron. Note that the kinetic energy is half in magnitude of the potential energy, but the potential energy is negative due to attraction. Writing the allowed discrete energies as

$$E_n = -\frac{m_e q^4}{2(4\pi\epsilon_0)^2\hbar^2}\frac{1}{n^2}, \tag{3.8}$$

Bohr found the difference between two energy states as

$$\boxed{E_{n_2} - E_{n_1} = \underbrace{\frac{m_e q^4}{2(4\pi\epsilon_0)^2\hbar^2}}_{\text{Ry=13.6eV}}(\frac{1}{n_1^2} - \frac{1}{n_2^2}),} \tag{3.9}$$

which is exactly equal to the experimentally observed spectral lines of Hydrogen! The Rydberg energy was thus discovered to be this unique combination of the electron mass m_e, its charge q, and Planck's constant $\hbar$. Bohr asserted that the reason for the stability of the atom is that the classical theory of mechanics that was used to argue that the atom should be unstable is simply incorrect for the electron. The mechanics of the electron follow quantum rules, and the question of instability simply does not arise because the discrete orbits are stable. Exactly why they are stable (i.e., stationary states) was proven mathematically much later, and the exact quantization condition also ran into experimental difficulties in explaining the spectrum of helium and other gases. But Bohr had, in one stroke brought quantum theory to particles, or electrons.

Fig. 3.10 de Broglie proposed in his PhD thesis that particles with mass have wavelengths associated with their motion. He was awarded the Nobel Prize in Physics in 1929.

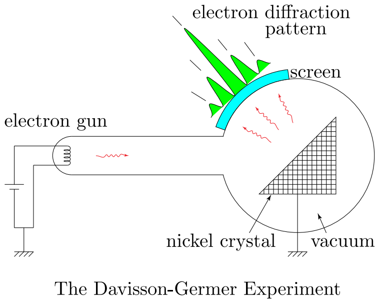


Fig. 3.11 The Davisson–Germer experiment that for the first time showed that electrons which are particles *diffract* from the surface of a nickel crystal. The periodically arranged atoms of the crystals diffract the electrons, proving conclusively their wave nature, and de Broglie's hypothesis that the wavelength λ and momentum p for both waves and particles are related by $\lambda = h/p$.

3.3 Wave-particle duality

One would think that Bohr, and Sommerfeld must have concluded from their own quantization condition $\oint pdx = nh$ that particles have wave nature because if for the electron $p = h/\lambda$, then one can immediately conclude that $n\lambda = \oint dx = 2\pi R$, meaning an allowed electron orbit is such that an integer number of electron wavelengths fit in *any* closed path length, not necessarily only the circular one. But such was the power of classical mechanics that it was difficult for even Bohr to accept that the relation $p = h/\lambda$, which was already found by Planck for photons or waves, could also apply to electrons. This deep connection – that there is a wavelength associated with particles with mass took another decade to be figured out. In 1924, de Broglie (Fig. 3.10) hypothesized in his PhD dissertation that classical "particles' with mass also have wavelengths associated with their motion. The wavelength is $\lambda = 2\pi\hbar/|\mathbf{p}|$, which is identical to $\mathbf{p} = \hbar\mathbf{k}$. How could it be proven?

The wavelength of visible light ($\sim 400 - 700$ nm) or photons is such that diffraction gratings (or slits) were available at that time. But electron wavelengths are much shorter[2]. Where might one find a diffrac-

---

[2] The electron wavelength is much shorter than a photon for the same energy as a photon. This is because an electron has a substantial momentum due to its mass compared to massless photons.

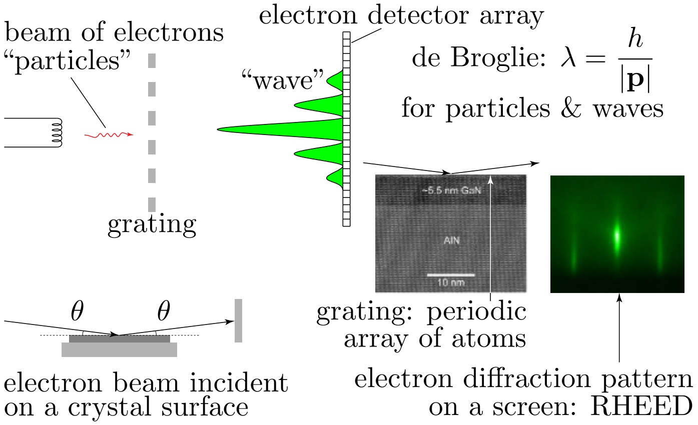


Fig. 3.12 Electrons behaving as waves. Electrons incident on a semiconductor crystal surface undergo diffraction from the periodic grating of atoms, producing characteristic patterns on a screen. ”Reflection high-energy electron diffraction” (RHEED) is a technique used routinely today for peering at the surface of a semiconductor crystal as it grows during molecular beam epitaxy (MBE). The technique a direct relic of the Davisson-Germer experiment that proves that electrons are waves.

tion grating that small? Max von Laue and co-workers in Munich had experimentally diffracted off X-rays from **crystals**, and discovered that the atoms in crystals are arranged periodically in space. In other words, a crystal is a *natural spectrometer* for X-rays, which are photons of wavelength of the order of interatomic distances of a few Å, where 1 Å=10^{-10} m. Crystals played a major role in the discovery of the wave-particle duality (and will play a major role in this book!). Elsassaer proposed using a crystal as a diffraction grating not for photons, but for electrons. A few years after de Broglie’s hypothesis (and unaware of it!), Davisson and Germer at the Bell Laboratories were shooting electrons in a vacuum chamber on the surface of crystalline nickel (Fig. 3.11)[3]. They observed diffraction patterns of electrons, which was mysterious to them, but as soon as they learnt of de Broglie’s hypothesis, they realized that they had unknowingly proven his hypothesis to be *quantitatively* correct. All particles had now acquired a wavelength! Fig. 3.12 completes the duality between the corresponding experiment with light (photons) that was discussed in Fig. 3.3.

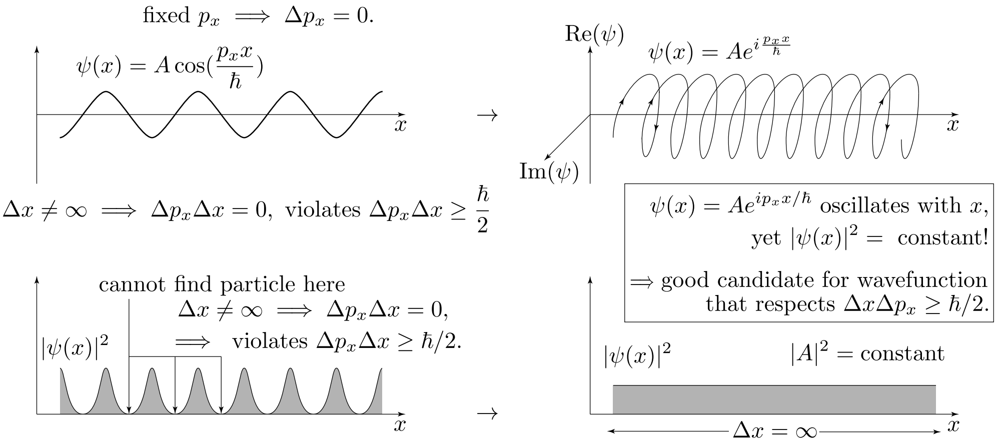
The Davisson–Germer experiment challenged the understanding of the motion or ”mechanics” of particles, which was based on Newton’s classical mechanics. In classical mechanics, the central question in every problem is the following: a particle of mass m has location x and momentum p now. If a force F acts on it, what are (x', p') later? Newton’s law $F = m\frac{d^2x}{dt^2}$ gives us the answer. The answer is deterministic, the particle’s future fate is completely determined from its present, and the force that acts on it. This is no longer correct if the particle has wave-like nature. The wave-particle duality is *the central fabric* of quantum mechanics. It leads to the idea of a wavefunction (Fig. 3.13).

Fig. 3.14 Joseph Fourier, the French mathematician and physicist introduced the powerful concept of Fourier series, which expresses any function as a sum of oscillating functions. Fourier’s ideas are central to the development of quantum mechanics that is consistent with the wave-particle duality. Fourier was a good friend of Napoleon.

3.4 The wavefunction

If a particle has a wavelength λ, what is its location x? A wave is an extended quantity. If a measurement of the particle’s location is per-

[3] The story goes that there was an accident and a leak developed in the vacuum apparatus sketched above. Because of the leak the surface of the nickel oxidized and lost its shine. They heated and melted the metal, and upon cooling it re-crystallized, and to their surprise, the electron diffraction spots showed up!


fixed $p_x \implies \Delta p_x = 0.$
$\psi(x) = A\cos(\frac{p_x x}{\hbar})$
x
$\rightarrow$
$\mathrm{Re}(\psi)$
$\psi(x) = Ae^{i\frac{p_x x}{\hbar}}$
x
$\mathrm{Im}(\psi)$
$\Delta x \neq \infty \implies \Delta p_x \Delta x = 0,$ violates $\Delta p_x \Delta x \geq \frac{\hbar}{2}$
$\psi(x) = Ae^{ip_x x/\hbar}$ oscillates with x, yet $|\psi(x)|^2 =$ constant!
$\Rightarrow$ good candidate for wavefunction that respects $\Delta x \Delta p_x \geq \hbar/2$.
cannot find particle here
$\Delta x \neq \infty \implies \Delta p_x \Delta x = 0,$
$\implies$ violates $\Delta p_x \Delta x \geq \hbar/2.$
$|\psi(x)|^2$
x
$\rightarrow$
$|\psi(x)|^2$
$|A|^2 =$ constant
$\Delta x = \infty$
x

- The state of the free quantum particle cannot be represented by independent numbers (x, p_x)
- We need a function whose amplitude oscillates in space, yet its magnitude never goes to zero.
- The complex exponential $e^{ik_x x}$ satisfies these requirements, and respects the uncertainty relation.

Fig. 3.13 Birth of the wavefunction to account for the wave-particle duality.

formed, it may materialize at location x_0. But repeated measurements of the same state will yield an average value $\langle x \rangle = x_0 + \Delta x$. Separate measurements of the momentum of the particle prepared in the same state will yield the average value $\langle p \rangle = p_0 + \Delta p$. The "uncertainty" relation states that $\Delta x \Delta p \geq \hbar/2$. This is a strictly mathematical consequence of representing a particle by a wave.

Because the *numbers* (x, p) of a particle cannot be determined with infinite accuracy simultaneously, one has to let go of the classical picture. How must one then capture the mechanics of a particle? Any mathematical structure used to represent the particle's state must contain information about its location x *and* its momentum p, since they are forever intertwined by the wave-particle duality. One is then forced to use a *function*, not a number. This function of a quantum object is denoted by ψ, and is called the **wavefunction**.

Fig. 3.15 Max Born introduced the probabilistic representation of the quantum wavefunction. With his student Heisenberg, he discovered matrix mechanics. He influenced and guided several young scientists who made contributions to quantum theory. Born was awarded the Nobel Prize in 1954.

A first attempt at constructing a wavefunction for a quantum particle with a momentum p is $\psi(x) = A\cos(px/\hbar)$. This guess is borrowed from the classical representation of waves in electromagnetism, and in fluid dynamics. This wavefunction can potentially represent a quantum particle of a definite momentum p if it satisfies the rule of wave-particle duality. Max Born (Fig. 3.15) provided the statistical interpretation of the wavefunction by demanding that $|\psi|^2$ be the **probability density**, and $\int |\psi|^2 dx = 1$. In this interpretation, $|\psi(x)|^2 dx$ is the probability that a measurement of the particle's location will find the particle in the location $(x, x+dx)$. $|\psi|^2$ is analogous to the *intensity* of an electromagnetic wave, and ψ to the electric field *amplitude*.

It is clear that $|\psi(x)|^2 = |A|^2 \cos^2(px/\hbar)$ assigns specific probabilities of the location of the particle, going to zero at certain points on the x-axis. Since the momentum p is definite, $\Delta p = 0$, and the uncertainty relation $\Delta x \Delta p \geq \frac{\hbar}{2}$ requires that $\Delta x \to \infty$, meaning the location of the particle must be equally probable at all points in space. Thus we reject the attempted wavefunction $\psi(x) = A\cos(\frac{px}{\hbar})$, because it is inconsistent with the uncertainty principle: it violates the wave-particle duality.

The simplest wavefunction consistent with the wave-particle duality picture is $\psi_p(x) = Ae^{ipx/\hbar}$, where A is a complex number constant. The complex exponential respects the wave-nature of the particle by providing a periodic variation in x, yet its probability density never goes to zero: the probability (density) is $|\psi_p(x)|^2 = |A|^2$, equal at all x. Thus, complex numbers are *inevitable* in the construction of the wavefunction representing a particle. In Fig. 3.17 it is shown that the best we can do to represent a particle at a particular location on the x-axis while respecting its wave nature is to create a **wavepacket** by creating a linear combination of a bunch of momentum values: $\psi(x) = \sum_p A_p e^{i\frac{px}{\hbar}}$, where A_p is the amplitude of each momentum component. If we try to localize the quantum particle tightly in real space x, Δx is small, but this wavepacket requires a larger number of p weights in the momentum space, and Δp becomes large. All linear superpositions of the form shown here are allowed in quantum mechanics. All the physical information of the quantum state is contained in the wavefunction $\psi(x)$ in real space, or equivalently, its representation in the momentum space, which is the set of coefficients $(..., A_{p1}, A_{p2}, ...)$. To extract any physical information from the wavefunction, we must apply the corresponding physical **operators** on the wavefunction.

Fig. 3.16 Werner Heisenberg discovered matrix mechanics with Max Born, and is an original founder of quantum theory. He was awarded the Nobel Prize in 1932. He is better known in pop media for his uncertainty principle.

3.5 Operators

Every physical observable in quantum mechanics is represented by an operator. When the operator "acts" on the wavefunction of the particle, it extracts the value of the observable. For example, the momentum operator is $\hat{p} = -i\hbar\partial/\partial x$, and for states of definite momentum $\hat{p}\psi_p(x) = (\hbar k)\psi_p(x)$. We note that $(x\hat{p} - \hat{p}x)f(x) = i\hbar f(x)$ for *any* function $f(x)$. The presence of the function in this equation is superfluous, and thus one gets the identity

$$x\hat{p} - \hat{p}x = [x, \hat{p}] = i\hbar. \tag{3.10}$$

The square brackets define a **commutation relation**. The space and momentum operators do not commute[4]. This sort of non-commuting mathematical structure of x and p was first uncovered by Heisenberg[5]. He was uncomfortable with the Bohr orbit picture, as it offered no explanation of why the coordinates (x, p) should follow an arbitrary quantization condition. Since experimentally electron orbits were not observed, but only the sharp spectral lines, Heisenberg was trying to

---

[4] Objects a and b commute if $ab = ba$ and do not commute if $ab \neq ba$. Clearly if a and b are real or complex numbers, they commute. But if they are matrices, or if one or both are differential operators such as $a = \frac{d}{dx}$ and $b = x^2$, they do not necessarily commute, since $\frac{d}{dx}(x^2 f(x)) \neq x^2 \frac{d}{dx} f(x)$ for an arbitrary function $f(x)$.

[5] "It was about three o' clock at night when the final result of the calculation lay before me. At first I was deeply shaken. I was so excited that I could not think of sleep. So I left the house and awaited the sunrise on the top of a rock." – Heisenberg.

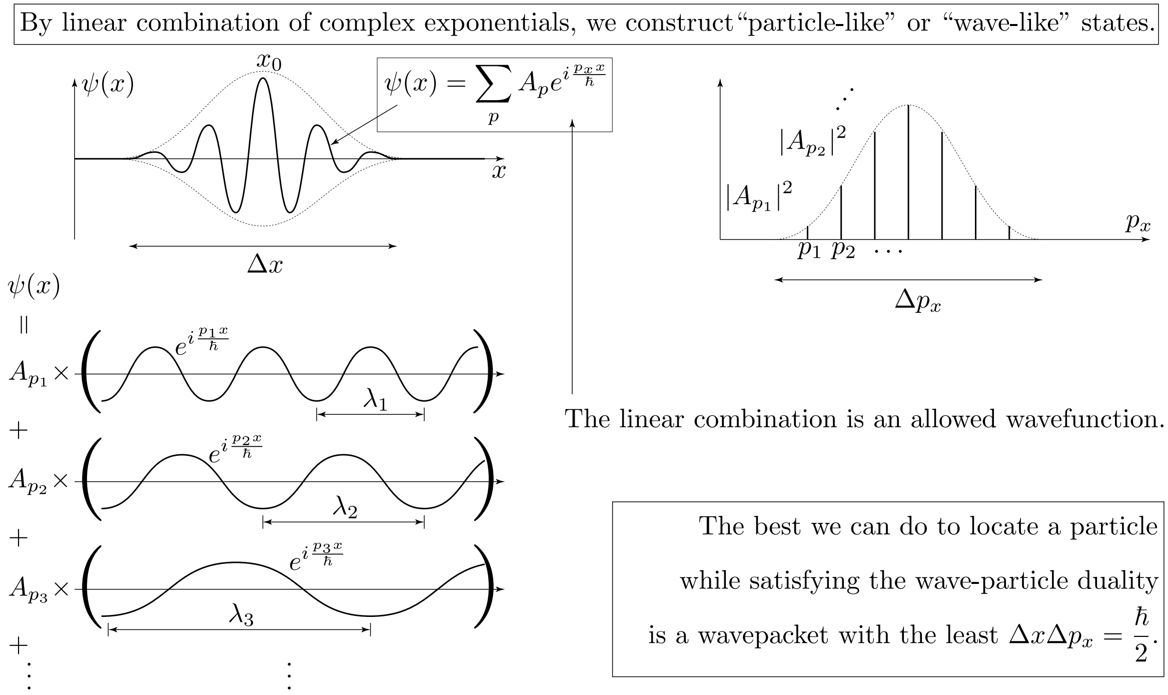
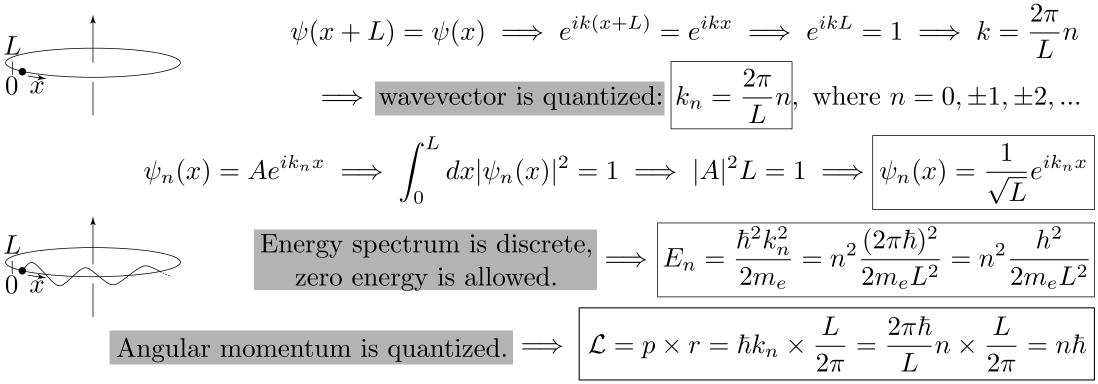

By linear combination of complex exponentials, we construct "particle-like" or "wave-like" states.
$\psi(x)$
x_0
$\psi(x) = \sum_p A_p e^{i\frac{p_x x}{\hbar}}$
x
Δx
$|A_{p_2}|^2$
$|A_{p_1}|^2$
p_x
p_1 p_2 $\cdots$
Δp_x
$\psi(x)$
$\|$
$A_{p_1}\times$
$e^{i\frac{p_1 x}{\hbar}}$
λ_1
$+$
$A_{p_2}\times$
$e^{i\frac{p_2 x}{\hbar}}$
λ_2
$+$
$A_{p_3}\times$
$e^{i\frac{p_3 x}{\hbar}}$
λ_3
$+$
The linear combination is an allowed wavefunction.
The best we can do to locate a particle while satisfying the wave-particle duality is a wavepacket with the least $\Delta x \Delta p_x = \frac{\hbar}{2}$.

Fig. 3.17 The superposition principle allows us to create wavefunctions that can represent "wave-like" or "particle-like" states with equal ease. Wave-like states have large Δx and small Δp, and particle-like states have small Δx and large Δp. All the while, they satisfy the uncertainty principle $\Delta x \Delta p \geq \hbar/2$.

construct a mathematically rigorous basis for the spectral lines. What he realized was that if x and p were *not numbers*, but objects that *did not commute*, then the differences between allowed energy levels could be *discrete* and explain the spectral lines. His mentor Max Born made the connection immediately that what Heisenberg had uncovered was that x and p are not numbers but *matrices*. This is because matrices do not necessarily commute. The first mathematically consistent formulation of quantum mechanics, due to Heisenberg, is thus called **matrix mechanics**. Matrices can be loosely considered to be "scars" of numbers, meaning the sharp numbers x and p are smeared out into a set of numbers arranged in the matrices $[x]$ and $[p]$.

In classical mechanics, $xp - px = [x, p] = 0$. Quantum mechanics elevates the "status" of x and p to those of mathematical operators (or equivalently, matrices), preventing them from commuting. This is referred to as the "first quantization" from classical to quantum mechanics. In this scheme, the dynamical variables (x, p) that were scalars in classical mechanics are promoted to operators, and the wavefunction ψ is a scalar. If the number of particles is not conserved, then one needs to go one step further, and elevate the status of the wavefunction $\psi \rightarrow \hat{\psi}$ too, which is called "second quantization".



$$\psi(x+L)=\psi(x) \implies e^{ik(x+L)}=e^{ikx} \implies e^{ikL}=1 \implies k=\frac{2\pi}{L}n$$

$$\implies \text{wavevector is quantized: } \boxed{k_n=\frac{2\pi}{L}n}, \text{ where } n=0,\pm1,\pm2,...$$

$$\psi_n(x)=Ae^{ik_nx} \implies \int_0^L dx|\psi_n(x)|^2=1 \implies |A|^2L=1 \implies \boxed{\psi_n(x)=\frac{1}{\sqrt{L}}e^{ik_nx}}$$

Energy spectrum is discrete,
zero energy is allowed.

$$\implies \boxed{E_n=\frac{\hbar^2k_n^2}{2m_e}=n^2\frac{(2\pi\hbar)^2}{2m_eL^2}=n^2\frac{h^2}{2m_eL^2}}$$

Angular momentum is quantized.

$$\implies \boxed{\mathcal{L}=p\times r=\hbar k_n\times\frac{L}{2\pi}=\frac{2\pi\hbar}{L}n\times\frac{L}{2\pi}=n\hbar}$$

Fig. 3.18 Quantum mechanics of the particle on a ring as an illustrative example of the effects of quantum confinement.

3.6 States of definite momentum and location

The wavefunction $\psi_p(x) = Ae^{ipx/\hbar}$ is a state of definite momentum since it is an **eigenstate** of the momentum operator $\hat{p}\psi_p(x) = p\psi_p(x)$. As a general rule, when an operator $\hat{O}$ of a physically observable quantity O acts on a quantum state ψ_O that is in a state of definite O, it extracts the *number* O from the wavefunction: $\hat{O}\psi_O = O\psi_O$. ψ_O is then called an **eigenfunction** of operator $\hat{O}$, and the number O is the **eigenvalue**. In the example discussed in this paragraph, $O = p$ is the eigenvalue of $\hat{O} = \hat{p} = -i\hbar\frac{\partial}{\partial x}$, and $Ae^{ipx/\hbar}$ is the corresponding eigenfunction[6].

One may demand the location of the particle to be limited to a finite length L as indicated in Fig. 3.18. This may be achieved by putting an electron on a ring of circumference L. A trial wavefunction for this particle is $\psi(x) = Ae^{ikx}$, which yields upon normalization $\int_0^L |\psi(x)|^2 dx = 1 \implies A = 1/\sqrt{L}$. The wavefunction must also satisfy the relation $\psi_p(x+L) = \psi_p(x)$ to be single-valued. This leads to $e^{ikL} = 1 = e^{i2\pi\times n}$, and $k_n = n \times (2\pi/L)$. Here $n = 0, \pm1, \pm2, ...$ The linear **momentum of the particle is then quantized**, allowing only discrete values. Since $L = 2\pi R$ where R is the radius of the ring, $k_nL = 2\pi n \to pR = n\hbar$, showing that the **angular momentum is quantized** to values[7] $0, \pm\hbar, \pm2\hbar, ...$ This indeed is the *quantum* of quantum mechanics! One may then index the wavefunctions of definite linear momentum by writing $\psi_n(x)$. Expressing states of definite momentum in terms of states of definite location similarly yields

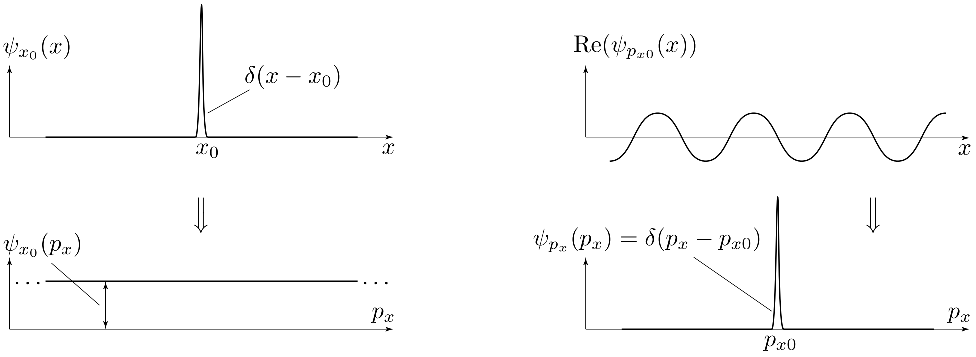
$$\psi_n(x) = \frac{1}{\sqrt{L}}e^{ik_nx}. \tag{3.11}$$

The set of wave functions $[...\psi_{-2}(x), \psi_{-1}(x), \psi_0(x), \psi_1(x), \psi_2(x), ...] = [\psi_n(x)]$ are special. We note that $\int_0^L dx\psi_m^\star(x)\psi_n(x) = \delta_{nm}$, i.e., the functions are orthogonal, and δ_{nm} is the Kronecker–delta. *Any* general wavefunction representing the particle $\psi(x)$ can be expressed as

---

[6] A possible confusion to avoid because of the limited symbols in English language here: the operator $\hat{O}$ and number O should not be confused with the number zero: 0.

[7] The angular momentum is a vector $\mathcal{L} = \mathbf{r} \times \mathbf{p}$ analogous to classical mechanics.



state of definite location

state of definite momentum

Fig. 3.19 States of definite location and states of definite momentum.

a linear combination of this set. This is the principle of superposition, and a basic mathematical result from Fourier's (Fig. 3.14) theory. Thus the quantum mechanical state of a particle may be represented as $\psi(x) = \sum_n A_n \psi_n(x)$. Clearly, $A_n = \int dx \psi_n^\star(x)\psi(x)$. Every wavefunction constructed in this fashion represents a permitted state of the particle[8], as long as $\sum_n |A_n|^2 = 1$.

It is useful here to draw an analogy to the decomposition of a vector into specific coordinates. In Fig. 3.20, the "hybrid" state function $\psi(x)$ is pictured as a vector $|\psi\rangle$ in an abstract space. The definite momentum wavefunctions $\psi_n(x)$ are pictured as the "coordinate" vectors $|n\rangle$ in that space of vectors. This set of vectors is called the basis. Since there are an infinite set of integers $n = 0, \pm 1, \pm 2, ...$, the vector space is infinite dimensional. It is called the Hilbert space. One may then consider the coefficients A_n as the length of the projections of the state on the basis states. The abstract picture allows great economy of expression by writing $|\psi\rangle = \sum_n A_n |n\rangle$. The orthogonality of the basis states is $\langle m|n\rangle = \delta_{mn}$, and thus $A_n = \langle n|\psi\rangle$. Then it is evident that $|\psi\rangle = \sum_n \langle n|\psi\rangle |n\rangle = \sum_n |n\rangle\langle n|\psi\rangle$, and $\sum_n |n\rangle\langle n| = 1$.

A vector may be decomposed in various basis coordinates. For example, a vector in 3D real space may be decomposed into cartesian, spherical, or cylindrical coordinate systems. Similarly, the choice of basis states of definite momentum is not unique. The wavefunctions for states of definite location are those functions that satisfy $x\psi_{x_0}(x) = x_0\psi_{x_0}(x)$, which lets us identify $\psi_{x_0}(x) = \delta(x - x_0)$. Here $\delta(...)$ is the Dirac–delta function, sharply peaked at $x = x_0$ (Fig. 3.19). It is instructive to expand the states of definite location in the basis of the states of definite momentum. From the uncertainty relation, we expect a state of definite location to contain many momenta. The expansion yields $A_n = \int_{-\infty}^{+\infty} dx/(2\pi/L) \times (e^{ik_n x}/\sqrt{L})\delta(x - x_0) = e^{ik_n x_0}/\sqrt{L}$, whereby $|A_n|^2 = 1/L$. Thus, the state of definite location x_0 is constructed of an infinite number of states of definite momentum $n = 0, \pm 1, \pm 2, ...$, each with equal probability $1/L$.

---

[8] Fig. 3.19 shows states of definite location and momentum. A general quantum state $\psi(x)$ can always be written as a linear superposition of states of definite location, or as a linear superposition of states of definite momenta.

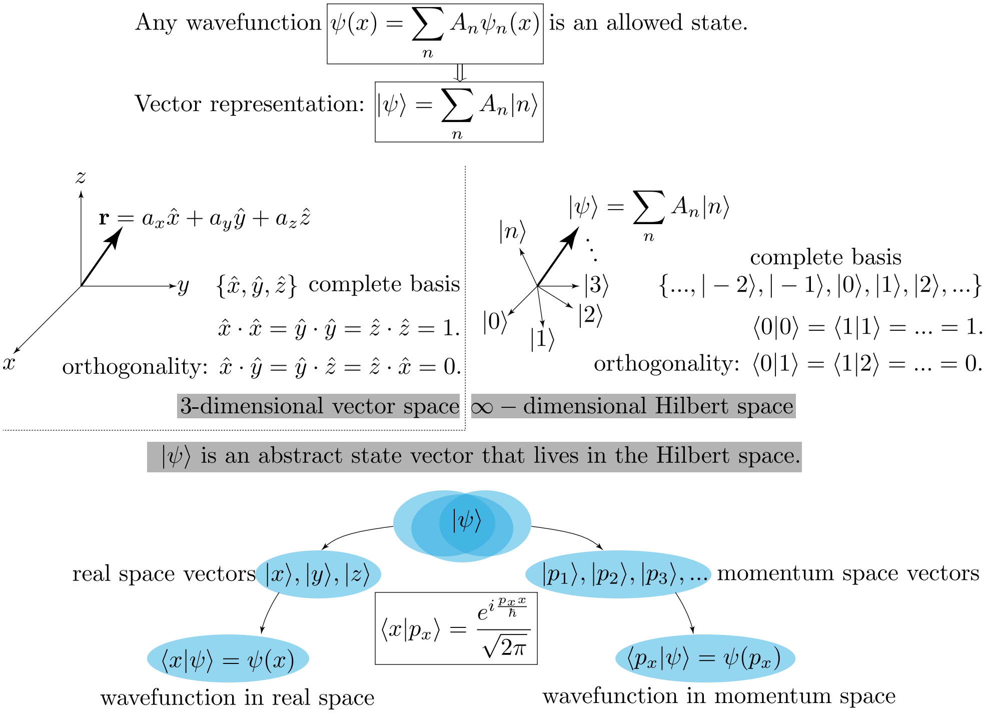

Any wavefunction $\psi(x) = \sum_n A_n \psi_n(x)$ is an allowed state.
Vector representation: $|\psi\rangle = \sum_n A_n |n\rangle$
z
y
x
$\mathbf{r} = a_x\hat{x} + a_y\hat{y} + a_z\hat{z}$
$\{\hat{x}, \hat{y}, \hat{z}\}$ complete basis
$\hat{x}\cdot\hat{x} = \hat{y}\cdot\hat{y} = \hat{z}\cdot\hat{z} = 1.$
orthogonality: $\hat{x}\cdot\hat{y} = \hat{y}\cdot\hat{z} = \hat{z}\cdot\hat{x} = 0.$
3-dimensional vector space
$|\psi\rangle = \sum_n A_n |n\rangle$
$|n\rangle$
$|3\rangle$
$|2\rangle$
$|1\rangle$
$|0\rangle$
complete basis
$\{..., |-2\rangle, |-1\rangle, |0\rangle, |1\rangle, |2\rangle, ...\}$
$\langle 0|0\rangle = \langle 1|1\rangle = ... = 1.$
orthogonality: $\langle 0|1\rangle = \langle 1|2\rangle = ... = 0.$
∞ − dimensional Hilbert space
$|\psi\rangle$ is an abstract state vector that lives in the Hilbert space.
$|\psi\rangle$
real space vectors $|x\rangle, |y\rangle, |z\rangle$
$|p_1\rangle, |p_2\rangle, |p_3\rangle, ...$ momentum space vectors
$\langle x|p_x\rangle = \frac{e^{i\frac{p_x x}{\hbar}}}{\sqrt{2\pi}}$
$\langle x|\psi\rangle = \psi(x)$
wavefunction in real space
$\langle p_x|\psi\rangle = \psi(p_x)$
wavefunction in momentum space

Fig. 3.20 Vector spaces for quantum states: we can use results of linear algebra for quantum mechanics problems.

3.7 States of definite energy: The Schrödinger equation

States of definite energy $\psi_E(x)$ are special. Unlike the states of definite momentum or definite location, we cannot write down their general wavefunction without additional information. That is because the energy of a particle depends on its potential and kinetic components. In classical mechanics, the total energy is $\frac{p^2}{2m} + V(x)$, i.e., it is split between kinetic and potential energy components. Once x and p are known for a classical particle, the energy is completely defined. However, since x and p cannot be simultaneously defined for a quantum-mechanical particle with arbitrary accuracy, the energy must be obtained through operations performed on the wavefunction.

Fig. 3.21 Erwin Schrödinger introduced the wave equation for quantum mechanics. He was awarded the Nobel Prize in 1933.

Schrödinger (Fig. 3.21) provided the recipe to find the states of definite energy, and the equation is thus identified with his name. The **Schrödinger equation** is

$$\boxed{\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right]\psi_E(x) = E\psi_E(x).} \tag{3.12}$$

The solution of this eigenvalue equation for a potential $V(x)$ identifies the special wavefunctions $\psi_E(x)$. These wavefunctions represent states of definite energy. How do we ascertain the accuracy of the Schrödinger equation? The answer: **through experiments**.

As discussed earlier, Niels Bohr had used a heuristic model to explain the spectral lines that lacked mathematical rigor. The triumph of the Schrödinger equation was in explaining the precise mathematical structure of the electron states, and the resulting spectral lines. An electron orbiting a proton in a hydrogen atom sees the potential $V(r) = -\frac{q^2}{4\pi\epsilon_0 r}$. Schrödinger solved this equation (with help from a mathematician), and obtained energy eigenvalues $E_n = -\frac{13.6}{n^2}$ eV. Thus Bohr's semi-qualitative model was given a rigid mathematical basis by Schrödinger's equation. The equation also laid down the recipe for solving similar problems in most other situations we encounter. Just as the case for states of definite momentum or definite location, one may expand *any* state of a quantum particle in terms of the states of definite energy $\psi(x) = \sum_E A_E \psi_E(x)$, or equivalently $|\psi\rangle = \sum_E A_E |E\rangle$

So why do states of definite energy occupy a special position in applied quantum mechanics? That is because they also mathematically explain why the electron in the allowed energy states of any atom is **stable**, that is, it does not change in time. That becomes clear if we consider the time-dependent Schrödinger equation.

3.8 Time-dependent Schrödinger equation

Newton's law $F = dp/dt$ provides the prescription for determining the future (x', p') of a particle given its present (x, p). Schrödinger provided the quantum-mechanical equivalent (Fig. 3.22), through the time-dependent equation

$$\boxed{i\hbar \frac{\partial \Psi(x,t)}{\partial t} = \underbrace{[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)]}_{\hat{H}} \Psi(x,t).} \tag{3.13}$$

To track the time-evolution of quantum states, one must solve this equation and obtain the composite space-time wavefunction $\Psi(x,t)$. Then physical observables can be obtained by operating upon the wavefunction by the suitable operators. Let's look at a particular set of solution wavefunctions that allow the separation of the time and space variables, of the form $\Psi(x,t) = \chi(t)\psi(x)$. Inserting this back into the time-dependent Schrödinger equation and rearranging, we obtain[9]

$$i\hbar \frac{\dot{\chi}(t)}{\chi(t)} = \frac{\hat{H}\psi(x)}{\psi(x)} = E. \tag{3.14}$$

Now since $i\hbar \frac{\dot{\chi}(t)}{\chi(t)}$ does not depend on space, and $\frac{\hat{H}\psi(x)}{\psi(x)}$ does not depend on time, yet both are equal at all space and time, both must

[9] Here $\dot{\chi}(t) = \frac{d\chi(t)}{dt}$.

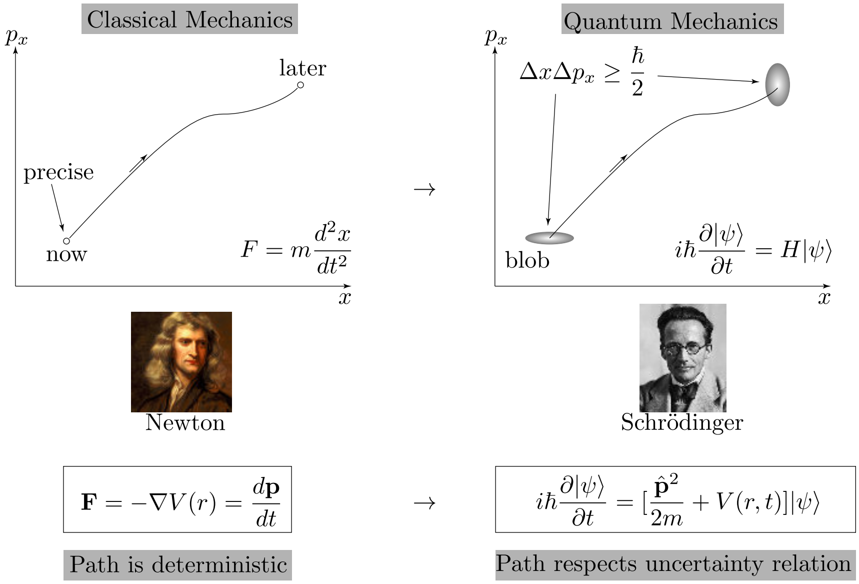

Classical Mechanics
p_x
later
precise
now
$F = m\frac{d^2x}{dt^2}$
x
$\rightarrow$
Quantum Mechanics
p_x
$\Delta x \Delta p_x \geq \frac{\hbar}{2}$
blob
$i\hbar\frac{\partial|\psi\rangle}{\partial t} = H|\psi\rangle$
x
Newton
Schrödinger
$\mathbf{F} = -\nabla V(r) = \frac{d\mathbf{p}}{dt}$
$\rightarrow$
$i\hbar\frac{\partial|\psi\rangle}{\partial t} = [\frac{\hat{\mathbf{p}}^2}{2m} + V(r,t)]|\psi\rangle$
Path is deterministic
Path respects uncertainty relation

Fig. 3.22 The dynamics of quantum states is governed by the time-dependent Schrödinger equation. Note that it looks like a hybrid of the classical energy and a wave equation, which is how it must be to account for the wave-particle duality.

be equal to a constant. The constant is called E, and clearly has dimensions of energy in joules. The right half of the equation lets us identify that $\hat{H}\psi_E(x) = E\psi_E(x)$ are states of definite energy. Then the left side dictates that the time dependence of these states is described by $\chi(t) = \chi(0)e^{-iEt/\hbar}$. Thus the particular set of solutions

$$\Psi_E(x,t) = \psi_E(x)e^{-i\frac{E}{\hbar}t} \tag{3.15}$$

now define the time evolution of the states. Here $\psi_E(x)$ are states of definite energy, as obtained by solving Equation 3.12, the time-independent Schrödinger equation[10].

3.9 Stationary states and time evolution

We note that $|\Psi_E(x,t)|^2 = |\psi_E(x)|^2$, that is, the probability of finding the particle at any x does *not* change with time. That means that a particle prepared in a state of definite energy will stay in that energy if there are no perturbations to $V(x)$. Its wavefunction does evolve in time as $\exp(-iEt/\hbar)$, but this evolution is "unitary" since its absolute value is unity. Notice the analogy with Newton's first law, which states that a particle at rest or moving with constant velocity will continue to do so unless acted upon by a force. The states of definite energy are therefore special since they do not evolve with time unless perturbed, and are called **stationary states**. They finally explain why the quantum mechanical states of electrons in atoms are stable. They are not really moving or orbiting in the classical sense, they are just there, spread in space as one diffuse object, with probability density $|\Psi_E(x,t)|^2 = |\psi_E(x)|^2$.

---

[10] An aspect of slowly varying, or adiabatic time evolution was missed by the original founders of quantum mechanics and only uncovered in the 1980's. These are embodied in the Berry phase, which is discussed in later chapters.

Any general allowed quantum state (not necessarily states of definite energy) may then be written as the linear superposition

$$\Psi(x,t) = \sum_E A_E \Psi_E(x,t) = \sum_E A_E \psi_E(x) e^{-i\frac{E}{\hbar}t}. \tag{3.16}$$

The states of definite energy form a convenient and often-used basis for expansion of general states of a particle. That is because they are stationary states – it is simpler if the basis states are fixed.

Consider the case where a hybrid state $\Psi(x,t)$ is prepared with components in two states $|E_1\rangle$ and $|E_2\rangle$. Then, the expansion is $\Psi(x,t) = A_{E_1}\psi_{E_1}(x)e^{-iE_1 t/\hbar} + A_{E_2}\psi_{E_2}(x)e^{-iE_2 t/\hbar}$. The probability density of this superposition state is

$$|\Psi(x,t)|^2 = |A_{E_1}\psi_{E_1}(x)e^{-iE_1 t/\hbar} + A_{E_2}\psi_{E_2}(x)e^{-iE_2 t/\hbar}|^2. \tag{3.17}$$

This is the single most important property of quantum states: we must **first add the amplitudes of the wavefunction, and then square** to find the probability density of superposed states. In other words, we add the amplitudes, not the intensities of the waves. Doing so yields:

$$\begin{aligned} |\Psi(x,t)|^2 = |A_{E_1}|^2|\psi_{E_1}(x)|^2 + |A_{E_2}|^2|\psi_{E_2}(x)|^2 + \quad & (3.18) \\ A_{E_1}A_{E_2}\psi_{E_1}(x)\psi_{E_2}(x)\cos\left(\frac{E_1 - E_2}{\hbar}t\right), \quad & (3.19) \end{aligned}$$

which **oscillates** with time with frequency $\omega_{12} = (E_1 - E_2)/\hbar$. Such two-level systems are being currently explored for making quantum-bits or qubits for a form of analog computation called quantum computation. If we had first squared and then added the intensities of the two states, their phase information would be lost, there will be no oscillations, and we will be back in the realm of classical mechanics.

All transport and optical phenomena involve time evolution. So most of the time in semiconductor physics we will work with the solutions of the time-dependent Schrödinger equation. The states of definite energy as a function of momentum $E(k)$ that form the **energy bandstructure** of the solid thus provide a most convenient basis for the analysis of electronic and optical phenomena of semiconductors.

The time evolution of the expectation value of an operator $\hat{O}$ is given by Ehrenfrest's theorem

$$\frac{d\langle\hat{O}\rangle}{dt} = -\frac{i}{\hbar}\langle[\hat{O},\hat{H}]\rangle, \tag{3.20}$$

where the operator itself is time-independent. By using $\hat{O} = \hat{p}$ and $\hat{H} = \frac{\hat{p}^2}{2m} + V(x)$, Ehrenfrest's theorem directly leads to Newton's second law. It forms the starting point for the density-matrix formulation of the time-evolution of quantum states.

3.10 Quantum current

In the physics of semiconductors and nanostructures, we will be deeply concerned with the flow of currents. A current is a measure of the flow of objects from one point in space to another. The flow of electric charges constitutes an electric current, leading to the notion of electrical conductivity. In this chapter we develop the recipe to understand current flow from a quantum-mechanical viewpoint. Since the physical state of particles in quantum mechanics is represented by its wavefunction $\Psi(x,t)$, the current must be obtained from the wavefunction.

Since $|\Psi(x,t)|^2 = \Psi^\star\Psi$ is the probability density, let's examine how it changes with time. We obtain

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = \Psi^\star\frac{\partial\Psi}{\partial t} + \frac{\partial\Psi^\star}{\partial t}\Psi, \tag{3.21}$$

where we use the time-dependent Schrödinger equation $i\hbar\partial\Psi/\partial t = (\hat{p}^2/2m + V)\Psi$ and its complex conjugate $-i\hbar\partial\Psi^\star/\partial t = (\hat{p}^2/2m + V)\Psi^\star$ to obtain for a time-independent potential $V(x)$

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = \Psi^\star\frac{(\hat{p}^2/2m + V)\Psi}{i\hbar} + \Psi\frac{(\hat{p}^2/2m + V)\Psi^\star}{-i\hbar}, \tag{3.22}$$

which simplifies to

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = \frac{1}{2mi\hbar}(\Psi^\star\hat{p}^2\Psi - \Psi\hat{p}^2\Psi^\star). \tag{3.23}$$

Since $\hat{p} = -i\hbar\nabla_{\mathbf{r}}$, we recognize the resulting equation

$$\frac{\partial|\Psi(x,t)|^2}{\partial t} = -\nabla_{\mathbf{r}}\cdot\left[\frac{1}{2m}(\Psi^\star\hat{p}\Psi - \Psi\hat{p}\Psi^\star)\right] \tag{3.24}$$

as the familiar **continuity equation** in disguise. A continuity equation is of the form $\partial\rho/\partial t = -\nabla_{\mathbf{r}}\cdot\mathbf{j}$, where ρ is the **particle number density** and $\mathbf{j}$ is the **particle current density**. This is illustrated in Fig. 3.23. We read off the quantum mechanical current density as

$$\boxed{\mathbf{j} = \frac{1}{2m}(\Psi^\star\hat{\mathbf{p}}\Psi - \Psi\hat{\mathbf{p}}\Psi^\star).} \tag{3.25}$$

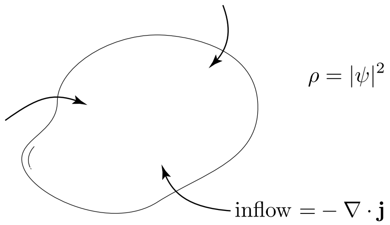


Fig. 3.23 The continuity equation $\frac{\partial\rho}{\partial t} = -\nabla\cdot\mathbf{j}$ states that the rate of increase of the particle density ρ at a point is equal to the net inflow of particles carried by the flux $\mathbf{j}$. The inflow is opposite to outflow, and is given by the divergence of the vector flux field.

This equation provides us the required recipe for calculating the probability density flow, or current flow directly from the quantum mechanical wavefunctions of states. We make a few observations. If Ψ is real, $\mathbf{j} = 0$. Since Ψ has dimension of $1/\sqrt{\text{Volume}}$, the dimension of $\mathbf{j}$ is per unit area per second. For 3D, volume is in m^3 and $\mathbf{j}$ is then in $1/(\text{m}^2\cdot\text{s})$. For 2D $\mathbf{j}$ is in $1/(\text{m}\cdot\text{s})$, and it is simply 1/s for 1D. We will use this concept of currents in greater detail in later chapters, and generalize it to charge, heat, or spin currents.

We also note that

$$\frac{d}{dt}\left(\int_{space} d^3r|\Psi|^2\right) = -\int_{space} d^3r\nabla\cdot\mathbf{j} = -\oint\mathbf{j}\cdot d\mathbf{S} = 0. \tag{3.26}$$

The conversion of the integral from volume to a closed surface uses Gauss' theorem. The value of the integral is zero because Ψ and consequently $\mathbf{j}$ goes to zero at infinity, and the equality must hold for all space. This equation is a statement of the indestructibility of the particle, which follows from $\int_{space} d^3r|\Psi|^2 = 1$. If the number of particles is *not* conserved, then one needs to add recombination ("annihilation") and generation ("creation") terms to the continuity equation. It then looks as $\partial\rho/\partial t = -\nabla \cdot \mathbf{j} + (G - R)$ where R and G are recombination and generation rates.

We also note that in the presence of a magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$ where $\mathbf{A}$ is the vector potential, the quantum-mechanical momentum operator $\hat{\mathbf{p}} \to \hat{\mathbf{p}} + q\mathbf{A}$ where q is the magnitude of the electron charge. This leads to an additional term in the expression of the current density

$$\mathbf{j} = \frac{1}{2m}(\Psi^\star \hat{\mathbf{p}}\Psi - \Psi\hat{\mathbf{p}}\Psi^\star) + \frac{q\mathbf{A}}{m}\Psi^\star\Psi. \tag{3.27}$$

The additional term depending on the magnetic vector potential $\mathbf{A}$ is needed to explain current flow in magnetic materials, magnetotransport properties, and superconductivity. It is discussed in Chapter 25.

If we want to determine the electric charge current, we realize that the current flux is actually of electrons that have wavefunction Ψ for which we have calculated the probability current flux $\mathbf{j}$. The charge q is dragged along by the electron. So to account for the flow of *charge*, the current density is simply $\mathbf{J} = q\mathbf{j}$, where q is the charge (in Coulombs) of the charge particle. If these charge particles are electrons, $q = 1.6 \times 10^{-19}$ C and free mass $m_e = 9.1 \times 10^{-31}$ kg. In the absence of a magnetic field, the electric current density is then given by

$$\boxed{\mathbf{J} = \frac{q}{2m_e}(\Psi^\star \hat{\mathbf{p}}\Psi - \Psi\hat{\mathbf{p}}\Psi^\star),} \tag{3.28}$$

which is now in A/m^2 for 3D, A/m for 2D, and A for 1D current flow, where A=C/s is the unit of current in Amperes. The current density is expressed in terms of the electron wavefunctions. We wish to connect the expression to classical Drude formulation of Chapter 2.

Consider free electrons in 1D with periodic boundary conditions between $x = (0, L)$. The wavefunction for a state $|k\rangle$ of definite energy $E(k)$ is $\Psi_E(x,t) = (1/\sqrt{L})e^{ikx}e^{-iE(k)t/\hbar}$. In the QM expression for current, the time evolution portion is not affected by the momentum operator, and therefore factors to 1. It is another illustration of the virtues of working with states of definite energy. The current carried by state $|k\rangle$ is then obtained as $J(k) = I(k) = q\hbar k/m_e L$. The current density and current are the same in 1D. The current $I(k) = q\hbar k/m_e L = qv(k)/L$ connects to the classical notion of current carried by a particle with velocity $v(k) = \hbar k/m_e$ traversing a distance L. Another way to picture the same current is to split it as $I = q \times v(k) \times n$, where $n = 1/L$ is the "volume density" of particles.

So we can find the current flow due to each allowed k-state for any quantum particle. Now let $f(k)$ be an occupation function that determines whether that k-state is occupied by a particle or not, and if it is, how many particles are sitting in it. To find the occupation function $f(k)$, we stumble upon one of the deepest mysteries of nature that was unearthed by quantum mechanics.

Fig. 3.24 Wolfgang Pauli discovered the exclusion principle for which he won the Nobel Prize in 1945. He introduced matrices for electron spin. He humorously referred to himself as the imaginary part of another notable physicist – Wolfgang Paul.

3.11 Fermions and bosons

The wave-particle duality requires quantum states to be represented mathematically by wavefunctions ψ, and $|\psi|^2$ represents the probability density. Together, these two seemingly simple requirements have a rather astonishing physical consequence for the properties of more than one quantum particle. Consider two quantum states $|a\rangle$ and $|b\rangle$ with real-space wavefunctions $\psi_a(x)$ and $\psi_b(x)$. What is the many-particle wavefunction when *two* particles are put in the two states? Let's label the locations of the two particles as x_1 and x_2. If the two particles are **distinguishable**, such as an electron and a proton, then the composite wavefunction may be written as the product of the single-particle wavefunctions (see Fig. 3.25)

$$\psi(x_1, x_2) = \psi_a(x_1)\psi_b(x_2). \tag{3.29}$$

But if the two particles are **indistinguishable**, such as two electrons, the wavefunction must satisfy further requirements[11]. Specifically, if we swap the locations of the two electrons $x_1 \leftrightarrow x_2$, the physical observables of the composite state must remain the same. This requirement dictates that the probability density must satisfy

$$P(x_2, x_1) = P(x_1, x_2) \to |\psi(x_2, x_1)|^2 = |\psi(x_1, x_2)|^2. \tag{3.30}$$

The product wavefunction of Equation 3.29 does not satisfy this requirement, because in general $|\psi_a(x_1)|^2|\psi_b(x_2)|^2 \neq |\psi_a(x_2)|^2|\psi_b(x_1)|^2$. So the simple product of the wavefunctions cannot represent indistinguishable particles. A symmetrized form, however does the job:

$$\psi(x_1, x_2) = \psi_a(x_1)\psi_b(x_2) + \psi_a(x_2)\psi_b(x_1) \tag{3.31}$$

because for this composite wavefunction,

$$\psi(x_2, x_1) = +\psi(x_1, x_2) \tag{3.32}$$

and the probability density does not change upon swapping. We also note that *both* particles may be at the same x since

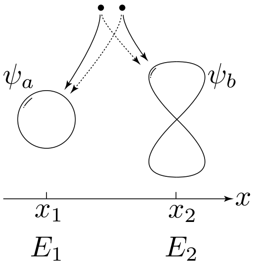
$$\psi(x_1, x_1) = +\psi(x_1, x_1) \tag{3.33}$$

is OK. Particles in nature that choose the "+" sign in Equation 3.32 are called **bosons**. Multiple bosons can occupy the same quantum state. What is very interesting is that this is not the only choice! The anti-symmetrized form

$$\psi(x_1, x_2) = \psi_a(x_1)\psi_b(x_2) - \psi_a(x_2)\psi_b(x_1) \tag{3.34}$$

---

[11] Because the total energy of two particles is the sum of the two, and the eigenfunctions evolve with time as $\Psi_E(x,t) = \psi_E(x)e^{-\frac{i}{\hbar}Et}$, the combined wavefunction of a multi-particle system must be some form of a **product** of the single particle eigenfunctions so that the energies add in the exponent.



$$\text{Total energy for 2 particles: } E = E_1 + E_2$$

$$\implies \text{ time evolution } \sim e^{-i(\frac{E_1+E_2}{\hbar})t}.$$

$$\text{Since } i\hbar\frac{\partial\psi}{\partial t} = E\psi,\ \psi \sim \underbrace{\psi_1}_{\sim e^{-i\frac{E_1}{\hbar}t}} \times \underbrace{\psi_2}_{\sim e^{-i\frac{E_2}{\hbar}t}}$$

$$\text{electron \& proton: distinguishable}$$

$$\implies \boxed{\psi(x_1,x_2) = \psi_a(x_1)\psi_b(x_2)}$$

$$\text{electron \& electron: indistinguishable!}$$

$$x_1 \leftrightarrow x_2 \text{ must leave } |\psi|^2 \text{ unchanged.}$$

$$\implies \boxed{\psi(x_1,x_2) = \psi_a(x_1)\psi_b(x_2) \pm \psi_b(x_1)\psi_a(x_2)}$$

Fig. 3.25 Indistinguishable particles suffer an identity crisis when we try constructing a wavefunction for more than one particle!

leads to

$$\psi(x_2, x_1) = -\psi(x_1, x_2), \tag{3.35}$$

which is also permitted, since the probability density remains unaltered by the negative sign[12] upon swapping the particles. Particles that choose the "-" sign are called **fermions**. However, an attempt to put both fermions in the same location leads to

$$\psi(x_1, x_1) = -\psi(x_1, x_1) \implies \psi(x_1, x_1) = 0. \tag{3.36}$$

This is the famous **Pauli exclusion principle**. It states the simple result that two identical fermions (e.g. electrons) cannot be in the same quantum state. The exclusion principle is responsible for the existence of the periodic table of elements, and all chemical behavior of matter.

Fig. 3.26 shows how the two core concepts of quantum mechanics: the Schrödinger equation, and the exclusion principle together explain the existence of the elements and the periodic table shown in Fig. 3.27. The structure of an atom consists of the nucleus with protons and neutrons. Electrons orbit the nucleus experiencing a net potential $V(r)$. The Schrödinger equation provides the recipe for finding the allowed **states of definite energy** for the electrons. The allowed set of definite energy states in which the electron is bound to the nucleus are referred to as **orbitals**, and are **discrete**, with corresponding discrete energy eigenvalues. For example, for a fluorine atom, the allowed electron orbitals are $\psi_{1s}(r)$, $\psi_{2s}(r)$, $\psi_{2p}(r)$... with the corresponding energy eigenvalues E_{1s}, E_{2s}, E_{2p}... Note that the p orbital allows three eigenvalues of equal energy. The orbitals of 2s and 2p are said to be in the same **shell**, which is different from 1s, and 3s, 3p, and 3d. **Each orbital allows electrons of two opposite spins**.

Because the Fluorine atom has 9 electrons, the electrons must start filling from the lowest orbital and follow the exclusion principle. Fol-

[12] Most practitioners of quantum mechanics will tell you how much they love this negative sign!

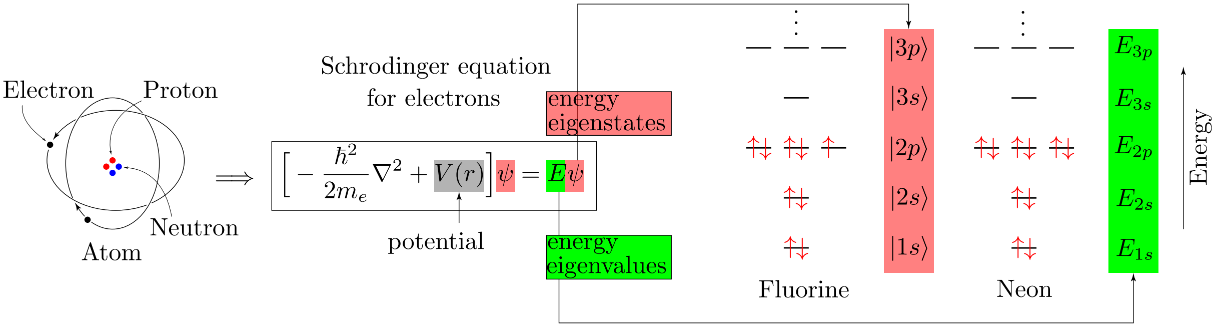
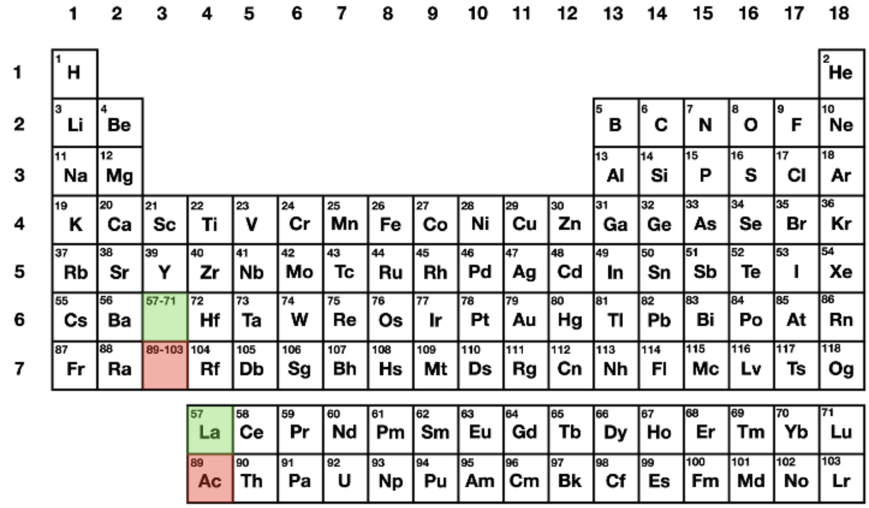


Fig. 3.26 The existence of the elements in the periodic table is explained by the exclusion principle, and Schrödinger's equation. The allowed electron states or orbitals, and energies are determined by the net potential experienced by it due to the combined influence of the nucleus and the other electrons by solving the Schrödinger equation. Filling these allowed orbitals with electrons according to the exclusion principle leads to a periodic occurrence of partially filled orbitals that make the elements chemically reactive - such as fluorine, and chemically inert, such as neon.

lowing this prescription, Fig. 3.26 shows that the 1s, 2s, and two of the 2p orbitals are completely filled, but one of the 2p orbitals has only one electron and can accept another. The next element in the periodic table is neon, which has 10 electrons. This exactly fills the 2p orbitals. This may seem like a minor detail, but the physical properties of fluorine and neon differ drastically due to this difference!

If all the orbitals of a shell are filled with electrons, we call it a **closed shell**. Atoms with 8 electrons in the outermost shell are inert gases. For neon, shells 1 and 2 are closed, and the atom is chemically **closed** too, meaning this atom will not accept, or give away electrons easily: there is a very high energy cost to do so. This makes neon chemically inert, and such elements are therefore called inert or **noble gases**. On the other hand, if fluorine could add just one more electron to its arsenal, it could become inert. This makes a fluorine atom extremely hungry for electrons – it will yank electrons away from any element that has one extra electron in a shell. Such atoms are H, Li, Na, ... or atoms in the 1st period of the periodic table. This is the reason for chemical bonding and the formation of molecules – if a hydrogen and a fluorine atom share one electron, both of them can effectively close their open shells. The hydrogen looks like helium, and the fluorine like neon in this molecule. The total energy of the hydrogen fluoride (HF) molecule is lower than the sum of the individual energies of the H and F atoms. This chemical reactivity of elements of group VII fueled by the hunger for that extra electron makes them bond with Group I metals, producing salts such as NaCl. This is the reason Group VII gases are called halogens, which literally means "salt-producing".

Pauli arrived at the exclusion principle in an effort to explain the sharp atomic spectra of helium, and many-electron atoms. The Bohr model, which was successful in explaining the hydrogen atom spectra, failed for atoms with two or more electrons. Enforcing the exclusion principle explained not just the spectra, but a whole host of other prop-



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57-71	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89-103	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
				57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
				89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Fig. 3.27 The periodic table of elements: all the (known) cards of the universe in one neat table!

erties of matter!

Many elements can continue lowering their energies by forming larger and larger molecules. A large molecule where the atoms are arranged in perfect periodic fashion is called a **crystal**. The elements to the left side of the periodic table have a few extra electrons in the outermost shells: they are mostly empty, and can lower their total energy by sharing electrons with nearby atoms. So they typically form crystals at room temperature – these crystals are called **metals**. The shared electrons in metals can move across the crystal easily, giving metals their high electrical conductivity. To the right of the periodic table are atoms such as O, N, Cl, and Fl that have mostly filled shells, these need a few electrons to lower their energies – which makes them form small molecules such as O_2 and N_2, and not condense into a crystal, but exist as a **gas** at room temperature. Near the middle of the periodic table, in group IV, the atoms C, Si, Ge tend to structurally form crystals like metals, but because they have exactly half-filled electronic shells, are electrically not conductive as the electrons are all used up in forming the chemical bond, with none left over for conduction. They are called **semiconductor crystals**, and are the class of materials that this book is primarily concerned with.

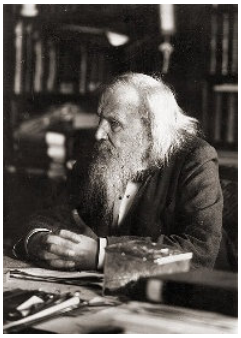


Fig. 3.28 Dmitri Mendeleev, the discoverer of the periodic table of elements. Based on the periodic occurrence of physical properties, Mendeleev *predicted* the existence of the then-unkown elements Ga and Ge, which he called eka-Aluminum and eka-Silicon possibly as a throwback to the ancient language of Sanskrit (eka=one), which uses periodically repeating speech sounds.

3.12 Fermion and boson statistics

Atoms and small molecules have a small number of electrons. But in very large molecules, or crystals as discussed in the last paragraph, there are a large number of energy orbitals available for electrons to occupy. Their distribution must follow a statistical process that minimizes the energy cost. For example, consider 10^{22} electrons in a solid, and 10^{23} orbitals – how do electrons choose which orbitals to occupy?

In the presence of large number of electrons, the Pauli exclusion principle leads to an occupation probability of quantum states that is obtained by the maximization of the thermodynamic entropy. The result was first derived by Fermi (Fig. 3.29) and Dirac for **non-interacting electrons**. It is derived in Chapter 4 from thermodynamic arguments coupled with the Pauli exclusion principle resulting from the wave-particle duality. The resulting distribution is called the Fermi–Dirac distribution

$$\boxed{f_{FD}(E) = \frac{1}{e^{\frac{E-\mu}{k_b T}} + 1},} \tag{3.37}$$

where μ is the chemical potential, which is a measure of the number of fermions, k_b the Boltzmann constant, and T the absolute temperature. The chemical potential μ grows with the number of particles, and depends on the temperature. At $T \to 0$ K, the Fermi–Dirac function assumes a unit step shape and the highest energy orbital that is occupied is called the Fermi energy E_F. So at $T = 0$ K, the chemical potential is equal to the Fermi energy $\mu(T = 0\text{ K}) = E_F$.

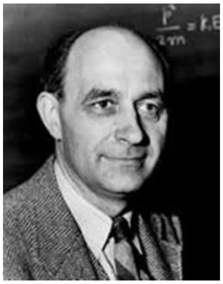


Fig. 3.29 Enrico Fermi made significant contributions to virtually all fields of physics. He was awarded the Nobel Prize in 1938.

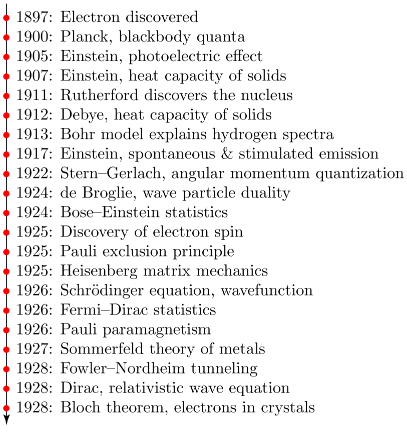
Note that the probability of a fermion occupying an orbital allowed by the Schrödinger equation **cannot exceed** 1, enforcing the exclusion principle. If a measurement to locate the fermion in a single orbital is made, the result will always be 0 or 1, not a fraction. But if a very large number of exact copies of the system are prepared, and the measurement to locate the fermion in a particular orbital is made on each system, the average occupation number will be given by the Fermi–Dirac distribution. In a single system if there are a large number of degenerate energy orbitals, then the occupation probability is the time-averaged value. This is the meaning of the distribution function.

Fig. 3.30 Satyendra Nath Bose, who discovered the statistics for photons. Particles that follow such statistics are called bosons.

The equivalent quantum-statistical result for bosons is

$$\boxed{f_{BE}(E) = \frac{1}{e^{\frac{E-\mu}{k_b T}} - 1},} \tag{3.38}$$

where μ is the chemical potential. The Bose–Einstein distribution allows values larger than 1. Dramatic effects such as the Bose–Einstein condensation (BEC), lasers, and the existence of superconductivity occurs when many bosons can co-exist in the same quantum state. The bosons can be composite particles, for example Cooper pairs in superconductors that are electron-phonon-electron quasiparticles where electrons are "glued" together by phonons.



1897: Electron discovered
1900: Planck, blackbody quanta
1905: Einstein, photoelectric effect
1907: Einstein, heat capacity of solids
1911: Rutherford discovers the nucleus
1912: Debye, heat capacity of solids
1913: Bohr model explains hydrogen spectra
1917: Einstein, spontaneous & stimulated emission
1922: Stern–Gerlach, angular momentum quantization
1924: de Broglie, wave particle duality
1924: Bose–Einstein statistics
1925: Discovery of electron spin
1925: Pauli exclusion principle
1925: Heisenberg matrix mechanics
1926: Schrödinger equation, wavefunction
1926: Fermi–Dirac statistics
1926: Pauli paramagnetism
1927: Sommerfeld theory of metals
1928: Fowler–Nordheim tunneling
1928: Dirac, relativistic wave equation
1928: Bloch theorem, electrons in crystals

Fig. 3.31 A timeline of the development of quantum mechanics and its applications to crystalline solids.

The Fermi–Dirac and Bose–Einstein statistics have a delightful history of development. The Bose–Einstein distribution was arrived at by analysis of the thermodynamic properties of a large number of light quanta, or photons, around 1924. Immediately after Pauli unearthed the exclusion principle, first Fermi, and then Dirac, arrived

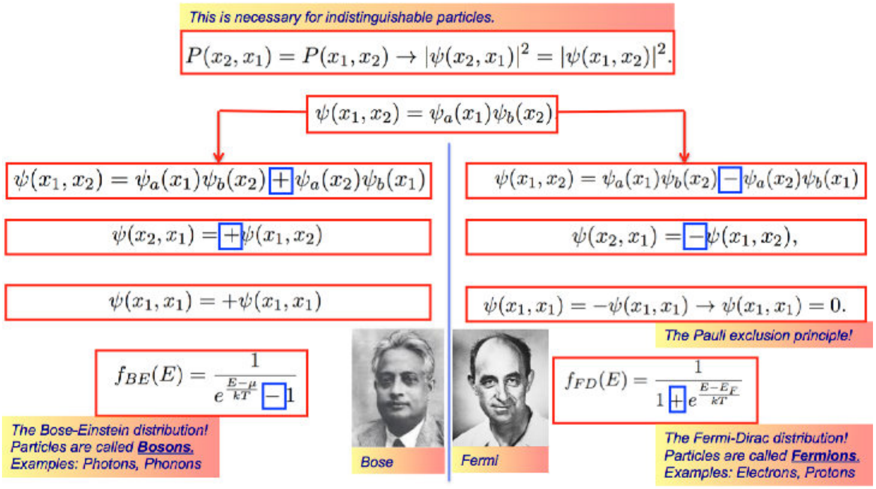

This is necessary for indistinguishable particles.
$P(x_2,x_1)=P(x_1,x_2)\rightarrow|\psi(x_2,x_1)|^2=|\psi(x_1,x_2)|^2$.
$\psi(x_1,x_2)=\psi_a(x_1)\psi_b(x_2)$
$\psi(x_1,x_2)=\psi_a(x_1)\psi_b(x_2)\boxed{+}\psi_a(x_2)\psi_b(x_1)$
$\psi(x_2,x_1)=\boxed{+}\psi(x_1,x_2)$
$\psi(x_1,x_1)=+\psi(x_1,x_1)$
$f_{BE}(E)=\frac{1}{e^{\frac{E-\mu}{kT}}\boxed{-}1}$
The Bose-Einstein distribution!
Particles are called **Bosons.**
Examples: Photons, Phonons
Bose
Fermi
$\psi(x_1,x_2)=\psi_a(x_1)\psi_b(x_2)\boxed{-}\psi_a(x_2)\psi_b(x_1)$
$\psi(x_2,x_1)=\boxed{-}\psi(x_1,x_2)$,
$\psi(x_1,x_1)=-\psi(x_1,x_1)\rightarrow\psi(x_1,x_1)=0$.
The Pauli exclusion principle!
$f_{FD}(E)=\frac{1}{1\boxed{+}e^{\frac{E-E_F}{kT}}}$
The Fermi-Dirac distribution!
Particles are called **Fermions.**
Examples: Electrons, Protons

Fig. 3.32 Indistinguishable particles can be of two types: Bosons, or Fermions. They have very different properties!

at the same distribution from separate angles. It was clear that the Bose–Einstein distribution applied to photons, and through models of vibrations of solids, was also known to apply to the thermodynamics of matter waves. But it was not immediately clear what the Fermi–Dirac distribution was to be applied to! Letters exchanged between Pauli and Heisenberg at that time show how Pauli was slowly convinced that the Fermi-Dirac distribution was the correct one for electrons and electronic phenomena in solid crystals. The first application of the Fermi-Dirac statistics was made by Pauli to explain the phenomena of **paramagnetism** of solids, which is a weak attraction non-magnetic solids feel towards magnets. Pauli's paramagnetism effectively launched solid-state physics and the quantum theory of solids. Arnold Sommerfeld was deeply impressed by Pauli's use of the Fermi-Dirac statistics to explain paramagnetism. He set out to fix the nearly three-decade old Drude model of metals. Sommerfeld had spectacular success in fixing the failings of the Drude model by simply changing the classical Maxwell Boltzmann distribution e^{-E/k_bT} to the new fully quantum Fermi–Dirac statistics of electrons. We will encounter this story in Chapter 5.

3.13 The Spin-statistics theorem

In addition to linear momentum $\mathbf{p}=\hbar\mathbf{k}$ and angular momentum $\mathbf{L}=\mathbf{r}\times\mathbf{p}$, electrons also possess an extra bit of spin angular momentum. In semiconductors, electron spin plays an important role in the electronic band structure. The net angular momentum of electron states is obtained by adding the various components of the angular momenta.

The exclusion principle is central to the spin-statistics theorem from relativistic quantum field-theory. Pauli proved the spin-statistics theorem rigorously in 1940, though its validity was well known by that time. It states that bosonic particles have integer spins, and fermonic particles have half-integer spins. That means bosons have spins $S = 0, \pm\hbar, \pm 2\hbar, ...$, and fermions have spins $S = \pm\hbar/2, \pm 3\hbar/2, ...$ Electrons have spin[13] $\pm\hbar/2$.

The fundamental dichotomy of particles in nature has received increasing attention the last three decades. Quasi-particle states have been observed (for example in the fractional quantum Hall effect) that behave neither like fermions nor bosons. Swapping the single-particle states for such quasi-particles leads to the accumulation of a phase factor:

$$\psi(x_2, x_1) = e^{i\phi}\psi(x_1, x_2). \tag{3.39}$$

These particles evidently satisfy the indistinguishability criteria, but accumulate a(ny) phase, leading to their name *anyons*. Exercise 3.10 explores some features of anyons further. Anyon states can exhibit a richer range of statistics than fermions and bosons. For anyons, commuting (or Abelian) statistics has similarity to fermions and bosons, but non-commuting (or non-Abelian) statistics does not have such an analog. Non-Abelian anyons are of current interest due to their proposed usage in topological quantum computation.

3.14 The Dirac equation and the birth of particles

Fig. 3.33 Paul Dirac unified quantum theory with Einstein's special relativity and discovered an equation that showed why the electron must have spin angular momentum. His equation gave birth to most of particle physics, and topological aspects of condensed matter physics. Shared the physics Nobel prize in 1933 with Schrödinger.

Dirac (Fig. 3.33) was not comfortable with Schrödinger's equation since it was not consistent with relativity, and did not predict the spin of electrons. He was able to reformulate (see Exercise 3.9) the quantum mechanics of electrons from Schrödinger's equation

$$i\hbar\frac{\partial|\psi\rangle}{\partial t} = [\frac{\hat{\mathbf{p}}^2}{2m} + V(\mathbf{r}, t)]|\psi\rangle \tag{3.40}$$

to the Dirac equation

$$i\hbar\frac{\partial|\psi\rangle}{\partial t} = [c\hat{\alpha}\cdot\hat{\mathbf{p}} + \beta mc^2 + V(\mathbf{r}, t)]|\psi\rangle \tag{3.41}$$

where c is the speed of light, and $\hat{\alpha}, \beta$ are the famous 4×4 Dirac matrices. The total energy for the free electron when $V(\mathbf{r}, t) = 0$ is given by $E = \sqrt{(pc)^2 + (mc^2)^2}$, which is consistent with special relativity. Before Dirac, the concept of a "particle" was not very clear. Dirac's assertion was to the effect: "a particle is the solution of my equation". Dirac's equation described the electron energy spectrum with more accuracy than Schrödinger's equation, and accounted for spin naturally. It also predicted the existence of negative energy states, or anti-electrons. This was the first prediction of **antimatter**. A few years after the prediction, such particles were discovered in cloud chambers by Carl Anderson (Fig. 3.34); these particles are called positrons. Electrons and positrons annihilate each other, emitting light of energy $\hbar\omega = 2m_0c^2$. The concept of a **hole** in semiconductors, which will be covered in quite some detail in this book has many analogies to a positron, but it is not quite the same. Positrons are used today in Positron Emission Tomography (PET) for medical imaging, and to also detect missing atoms (or vacancies) in semiconductor crystals.

---

[13] A more accurate statement is that when the spin angular momentum of the electron is *projected along any axis*, it can take the values $\pm\hbar/2$. This applies to all the earlier statements in this paragraph.

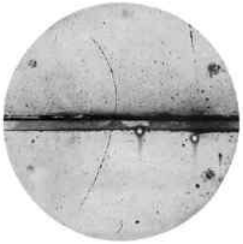


Fig. 3.34 The positron, or the anti-electron, was first anti-matter particle that was predicted to exist by the Dirac equation. It was discovered by Carl Anderson in cloud chamber tracks in 1934.

The philosophy of Dirac that "particles are solutions to equations" gave rise to the prediction of a number of new particles that have since been observed such as quarks, gluons, Higgs boson, etc... Majorana fermions fall under the category of predicted exotic particles, and there is intense interest in realizing such exotic states in matter for topological quantum computation. What was exotic yesterday will become commonplace tomorrow, so keep track of those "particles"!

3.15 Chapter summary section

In this chapter, we have learned:

- All objects with momentum p have a wavelength $\lambda = \frac{h}{p}$, where h is the Planck's constant. This is the de Broglie relation of the wave-particle duality. It holds for massless photons for which $p = \hbar\omega/c$, and also for particles with mass such as electrons for which $p = \frac{m_e v}{\sqrt{1-(\frac{v}{c})^2}} \approx m_e v$ for $v << c$.
- Because a wave is extended in space, the location x and the momentum p_x of a quantum particle such as an electron cannot be identified exactly; their uncertainties are related by the Heisenberg uncertainlty relation $\Delta x \Delta p_x \geq \frac{\hbar}{2}$, and the space and momentum coordinates do not commute any more: $xp_x - p_x x = i\hbar$.
- Since the classical mechanics coordinates (x, p) are now linked by the wave-particle duality, the state of a quantum system at a given time t is completely represented by a wavefunction $\psi(x, p, t)$ that has both (x, p) in it, or equivalently by a state vector $|\Psi\rangle = |\Psi(\mathbf{r}, t)\rangle$. For example, the wavefunction of a free electron moving in 1D is $\psi(x, p, t) = Ae^{i\frac{px}{\hbar}}e^{-i\frac{\frac{p^2}{2m_e}}{\hbar}t}$.
- For a physically observerable quantity A there is an operator $\hat{\mathbf{A}}$. The eigenvalues of $\hat{\mathbf{A}}$ are the possible results of the measurements of A, that is, denoting the eigenvalues of $\hat{\mathbf{A}}$ by a, $\hat{\mathbf{A}}|a\rangle = a|a\rangle$, and the probability of a measurement of A yielding the value a at time t is $|\langle a|\Psi(t)\rangle|^2$. The a's, which are the results of possible measurements, must be real. This implies that $\hat{\mathbf{A}}$ must be a linear hermitian operator.

- A measurement of $|\Psi\rangle$ that leads to an eigenvalue a_i leads the quantum mechanical system to *collapse* into the eigenstate $|\Psi_i\rangle$, which is the eigenstate corresponding to the eigenvalue a_i. So a measurement affects the state of the quantum system.
- Unlike classical mechanics where the total energy of a particle $\frac{p^2}{2m} + V(x)$ is completely defined by defining (x, p), the wave-particle duality forbids knowing (x, p) precisely at any time. Therefore, the energy has to be found by solving the time-independent Schrödinger equation $(\frac{\hat{p}^2}{2m} + V)\psi_E = E\psi_E$. The solutions of this wave equation ψ_E are the states of definite energy. The allowed energies can be sharp and discrete as in atoms and molecules, or continuous as for free electrons, or bands in solids.
- The time-evolution or dynamics of a classical system is governed by Newton's law $F = m\frac{d^2x}{dt^2}$. The time evolution of a quantum system is obtained by solving the time-dependent Schrödinger equation $i\hbar\frac{\partial\psi(x,t)}{\partial t} = (\frac{\hat{p}^2}{2m} + V)\psi(x,t)$. The states of definite energy of the time-independent Schrödinger equation are the stationary states of a quantum system, they do not change with time. Any linear combination of allowed quantum states is also an allowed quantum state.
- All quantum particles are either bosons, or fermions. An allowed quantum state can hold any number of bosons. A quantum state can only hold 0, or 1 fermions, which is the statement of the Pauli exclusion principle. Photons are bosons, and electrons are fermions.
- All quantum particles possess an internal angular momentum called spin. Fermions have half-integer spins $(\frac{1}{2}\hbar, \frac{3}{2}\hbar, ...)$, and bosons have integral spins $(0, \hbar, 2\hbar, 3\hbar, ...)$.
- In classical thermodynamics, the probability of occupation of a particle of energy E in equilibrium with a heat bath of temperature T is $e^{-\frac{E}{k_b T}}$. In quantum mechanics, a large number of bosons follow the Bose–Einstein statistics $f_{BE}(E) = \frac{1}{e^{\frac{E-\mu}{k_b T}} - 1}$, and a large number of fermions follow the Fermi–Dirac statistics $f_{FD}(E) = \frac{1}{e^{\frac{E-\mu}{k_b T}} + 1}$.

Further reading

With so many books written on quantum mechanics in the last 100 years, I will suggest the few books that I have personally had the pleasure to read critically and learn from over the years and would like to recommend.

For a physical and heuristic introduction, Volume 3 of **The Feynman Lectures on Physics** by Feynman, Leighton, and Sands still remains the book that says it as it is and is fun to read.

I found **Quantum Mechanics** by R. Shankar the most comprehensive to learn from, and the bits of humor sprinkled throughout help the reader to work through the chapters and exercises. **Modern Quantum Mechanics** by J. J. Sakurai uses surprisingly clear conceptual paths for most topics. Both these books have a collection of practice problems and can be used to learn the subject.

For engineers and applied physicists who are the true "quantum mechanics", I highly recommend **Quantum Mechanics** by H. Kroemer, who I was fortunate to have as my teacher and a mentor in graduate school. His book is the best complement to this textbook.

As your acquaintance with some of the counterintuitive concepts of quantum mechanics grows, and you want to learn more, I highly recommend **Lectures on Quantum Mechanics** by G. Baym which takes you to deeper into the subject in a set of very clear lecture notes.

The two classics that will serve as eternal resources for enthusiasts of the subject are **Quantum Mechanics**, Vol. 3 of Landau and Lifshitz's course of Theoretical Physics, and **The Principles of Quantum Mechanics** by P. Dirac. Both these books belong to anyone's desk who has an interest in learning the subject from the original masters.

Exercises

(3.1) Lord of the ring

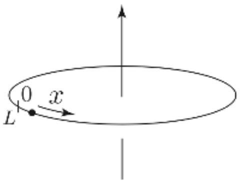


Fig. 3.35 Electron on a ring.

As discussed in this chapter, the allowed wavefunctions representing an electron on a circular ring (Fig. 3.35) of circumference L is $\psi_n(x) = \frac{1}{\sqrt{L}}e^{ik_n x}$, where $k_n = \frac{2\pi}{L}n$ are *quantized* because $n = 0, \pm1, \pm2, \ldots$. The *angular momentum* of a particle is defined as $\mathcal{L} = \mathbf{r} \times \mathbf{p}$, where $\mathbf{r}$ is the "radius" of the circle, and $\mathbf{p}$ is the linear momentum.

(a) Show that the angular momentum of an electron in state $\psi_n(x)$ is $\mathcal{L}_n = n\hbar$, where $\hbar = \frac{h}{2\pi}$ is the "reduced" Planck's constant. This implies that the angular momentum is *quantized* to values $0, \pm\hbar, \pm2\hbar, \ldots$. Compare the quantized angular momentum $\mathcal{L}_1$ for $n = +1$ with the classical angular momentum $\mathcal{L}_{cl}$ of a mass $m = 1$ kg being spun by a string of length $R = 1$ m with tangential velocity $v = 1$ m/s to appreciate how "nano" is the quantum of angular momentum.

(b) By balancing the classical centrifugal force and the electromagnetic Lorentz force, show that for an electron to be in the quantum state $\psi_n(x)$ on the ring, we need a *magnetic* field B_n such that the magnetic flux is $\Phi_n = B_n \cdot A = n \times \frac{h}{2q}$. Here A is the area of the ring, q is the electron charge and $h = 2\pi\hbar$. The quantity $\frac{h}{q}$ is known as the *quantum of magnetic flux*, and $\mu_B = \frac{q\hbar}{2m_e}$ is called the *Bohr magneton*. They have been measured experimentally in several semiconductors. Chapter 25 discusses these quantities in greater detail.

(c) Consider the quantum state obtained by the superposition $\psi(x) = a[\psi_{n=1}(x) + \psi_{n=-1}(x)]$ from the eigenstates of the electron on the ring. Normalize the state to find the constant a. You may need the result $\int_0^L \cos^2(\frac{2\pi}{L}x)dx = \frac{L}{2}$. Does this superposition state have a definite momentum?

(d) We derived that the quantum expression for current flux is $\mathbf{j} = \frac{1}{2m}(\psi^\star\hat{\mathbf{p}}\psi - \psi\hat{\mathbf{p}}\psi^\star)$, where $\hat{\mathbf{p}} = -i\hbar\nabla$ is the momentum operator, which takes the form $\hat{p}_x = -i\hbar\frac{\partial}{\partial x}$ for the particle on the ring. Show that even though the states $\psi_{n=1}(x)$ and $\psi_{n=-1}(x)$ carry *net* currents, their superposition state of part (c) does not. Explain.

(3.2) **Planck's dilemma**

In this problem, we re-derive Planck's famous formula for blackbody radiation, which gave birth to quantum mechanics. As usual, $k = 2\pi/\lambda$ is the wavevector, and $\omega = 2\pi\nu$ is the circular frequency in cycles/second.

(a) Show that the number of modes of a light wave that can fit in a cubic box of edge L between modes of wavevector k and $k + dk$ is $2 \cdot \frac{4\pi k^2 dk}{(\frac{2\pi}{L})^3}$. Since for light $\omega = ck$ where c is the speed of light, this implies that the density of optical modes, or density of states of light in a box of size $V = L^3$ is

$$D_\omega(\omega) = \frac{\omega^2 V}{\pi^2 c^3}. \tag{3.42}$$

(b) Show that if we take Planck's prescription that in any mode the energy is quantized to values $E = n\hbar\omega$ where $n = 0, 1, 2, ...$, and still assume classical thermodynamics and attribute a probability of occupation $e^{-\frac{E}{k_b T}}$ to each quanta, the energy allocated to a mode of energy $\hbar\omega$ is not the classical equipartition value $k_b T$ any more, but

$$\langle E \rangle = \hbar\omega \cdot \underbrace{\frac{1}{e^{\frac{\hbar\omega}{k_b T}} - 1}}_{n(\omega)}, \tag{3.43}$$

where the "average" number of photons in mode ω and energy $\hbar\omega$ is identified as $n(\omega)$.

(c) Now get Planck's result for the spectral energy density per unit volume $u_\omega(\hbar\omega) = \frac{\mathcal{U}_\omega(\hbar\omega)}{V} = \frac{D_\omega(\omega)}{V} n(\omega)\hbar\omega$,

$$\boxed{u_\omega(\hbar\omega) = \frac{\hbar}{\pi^2 c^3} \frac{\omega^3}{e^{\frac{\hbar\omega}{k_b T}} - 1}.} \tag{3.44}$$

This is the famous Planck's formula of blackbody radiation that resolved the ultraviolet catastrophe (Fig. 3.2) and launched quantum mechanics.

(d) Show that in the classical limit, when the discreteness of light quanta is much smaller than the thermal energy $\hbar\omega << k_b T$, the classical equipartition of energy is recovered:

$$\langle E \rangle|_{\hbar\omega << k_b T} = \langle E \rangle_{cl} \approx k_b T, \tag{3.45}$$

which yields the classical Rayleigh–Jeans law for the low-energy or infrared part of the spectrum:

$$u_\omega(\hbar\omega) \approx \frac{\omega^2}{\pi^2 c^3} k_b T. \tag{3.46}$$

Note that Planck's constant $\hbar$ does not appear in the classical result.

(e) Make a plot of Planck's spectrum given by Equation 3.44 and the Rayleigh–Jeans result Equation 3.46 for temperatures T = 300, 3000, 4500, and 5800 K. Chapter 28 shows the importance of these temperatures and the Planck distribution for the operation and efficiency limits of semiconductor solar cells.

(3.3) **The photoelectric effect**

Like the blackbody spectrum, another experimental observation that was puzzling at the time was the photoelectric effect. The experiment is shown schematically in Fig. 3.36. Light of frequency ω is incident on a clean surface of a metal, and an electrode is connected to collect any electrons that are ejected from the metal. The kinetic energy E_{kin} of the ejected electrons is measured. When this experiment is performed, the measured kinetic energy of an individual electron is shown in Fig. 3.36 as a function of the frequency, or the energy of the light.

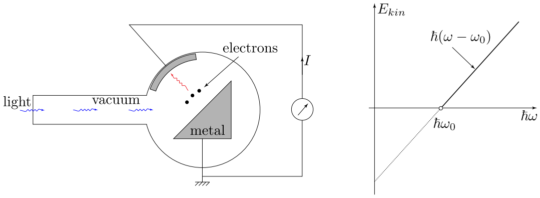


Fig. 3.36 The photoelectric effect.

The curious effect is that classically, one expects that increasing the intensity of the light beam at any energy would increase the kinetic energy of the ejected electrons. However, this is experimentally not the case. The kinetic energy of each ejected electron depends *only* on the *frequency* of the incident light beam, not its intensity.

Einstein provided an explanation of the photoelectric effect by combining Planck's idea of the quanta of light, with the recently discovered matter particle, the electron. His argument goes like this:

(a) If Planck's hypothesis is correct, then the light beam of frequency ω is composed of an integer number of photons, each of energy $\hbar\omega$. But electrons in a metal are bound to it with an energy $q\phi_M$, which is its workfunction, typically $\sim$ 4 eV. Now if the energy of the photon is lower than the workfunction, it is not possible for the photon to eject any electron at all, no matter how much its intensity. Is this argument reasonable – can you find any possible flaws and find why it is correct most of the time?

(b) Now for photons of energy larger than the workfunction, show that the kinetic energy of the ejected electrons is

$$E_{kin} = \hbar(\omega - \omega_0) \tag{3.47}$$

where $\hbar\omega_0 = q\phi_M$ is the work function of the metal. Argue why the kinetic energy of each electron cannot depend on the intensity of the light beam either! Show why a measurement of the kinetic energy of the electrons will yield a figure like what is sketched in Fig. 3.36, and how this measurement will simultaneously provide an experimental determination of the workfunction of the metal $q\phi_M$, and the Planck's constant $\hbar$.

(c) Show why the only quantity the intensity of the light beam can modify is the *number* of ejected electrons. This will lead to an increase in the current that flows in the collector circuit.

(d) The experimental confirmation of Einstein's prediction was achieved by Robert Millikan in 1914, and put quantum mechanics on firm foundation. Both Einstein and Millikan were awarded the Nobel prize for the photoelectric effect. Research Millikan's experiment, and its impact on future of vacuum tubes and electronics.

(e) Argue why the above mechanism is a method to convert light energy to electrical energy. This mechanism is used both for detecting photons in photomultiplier tubes (PMTs), and the concept underlies all solar cells.

(3.4) Bohr's orbits and the big bang

Bohr, together with Sommerfeld introduced the quantization condition $\oint pdx = nh$ to explain the stability of the atom, and the measured spectral lines of gases. Answer the following questions to develop an appreciation for the original Bohr orbit theory, or the "Old Quantum Theory" further.

(a) Calculate the magnitude of the Bohr radius (Equation 3.6) and the Rydberg energy (Equation 3.7) for the electron. Discuss why the Rydberg energy with $n = 1$ is the ionization energy of the electron in hydrogen.

(b) Re-calculate the effective Bohr radius $a_B^\star$ and the effective ionization energy $Ry^\star$ if the mass of the quantum particle is $m^\star = 0.1 m_e$ and the effective dielectric constant is $10\epsilon_0$. These two quantities are of significant interest in semiconductor physics to explain the phenomena of **doping**.

(c) By calculating the Coulomb attraction force between the electron and the proton, argue why and how according to classical mechanics, the hydrogen atom is unstable. If the electron and proton were *uncharged*, what would the radius of its classical orbit be?

(d) Calculate the wavelengths of spectral lines of hydrogen, and list those that fall in the visible range. Now when astronomers pointed their telescopes to distant galaxies and measured the spectra of light from them, they found the Hydrogen spectral lines again, except they find these lines are slightly *red-shifted* from the measurements performed on hydrogen gas on earth. This shift was used to prove that the universe is expanding, and started out from a big bang. Explore the magnitude of the red shifts, and how the age of the universe can be concluded from these shifts.

(e) Find out how the Bohr orbit theory failed to explain the spectrum of multi electron atoms such as helium, and how this discrepancy led to the unearthing of the Pauli exclusion principle.

(3.5) Electron waves and the quantum bumps

The Davisson–Germer experiment, discussed in Section 3.3, is that of an electron beam scattering off the surface of a crystal and producing diffraction patterns. It was the earliest experimental proof of the de Broglie relation $\lambda = h/p$ of wave-particle duality. In this problem, we calculate a few details of this experiment, and discuss its implications, and its many uses today.

(a) Calculate λ vs. the momentum p and energy E of electrons moving in vacuum. Make a log-log plot of the wavelength of electrons vs. their momentum and their energy in eV. Make the wavelength go over the range $10^{-15}\text{m} \leq \lambda \leq 10^{3}\text{m}$. What is the wavelength of the electron when its energy is $m_e c^2$, where c is the speed of light? This wavelength scale is called the Compton wavelength, which is the topic of Exercise 3.6.

(b) Einstein's theory of relativity forbids any particle with mass from going faster than the speed of light c. The de Broglie wave-particle duality relation for the electron of mass m_e consistent with relativity is $\lambda = \frac{h}{m_e v}\sqrt{1-(\frac{v}{c})^2}$. Calculate and plot the wavelength of the electron as a function of its velocity v. Superpose the relativistic result on one where $\lambda = \frac{h}{m_e v}$, and relativity is neglected. Around what velocities do you see significant deviations?

(c) The distance between atoms in a crystal is typically in the range of $a_0 \sim 0.5$ nm, or 5Å. Show that a diffraction pattern is obtained for electrons, show that their energy should be of the order $\frac{h^2}{2m_e a_0^2}$. Show that this is a few eV of energy, which is considered as a low energy. The technique of diffracting such electrons from the surface of crystals is used today to determine the crystal structure, and is called exactly what it is: Low Energy Electron Diffraction, or **LEED**.

(d) Historically, Elasasser had proposed the electron diffraction experiment, but a high vacuum was necessary to perform the experiment. Why is vacuum necessary, and what sort of vacuum levels (in Torr of pressure, with standard atmosphere being 760 Torr $\sim 10^5$ Pascal) were needed for the experiment? The reason Davisson and Germer succeeded in observing the diffraction patterns (they called it the "quantum bumps") is their access to a higher level of vacuum and high quality crystals compared to their counterparts.

(3.6) **The Compton effect**
One of the earliest experiments to show the particle nature of light was Compton scattering. Just as two particles (say balls) can collide and change their momenta, Compton showed in the early 1920s that when high energy X-ray photons scatter with electrons, the momentum of the photon chances, and as a result of the relation between the momentum and wavelength, the wavelength changes too. Note that this result was obtained *before* the de Broglie hypothesis of wave-particle duality, and likely influenced it heavily.

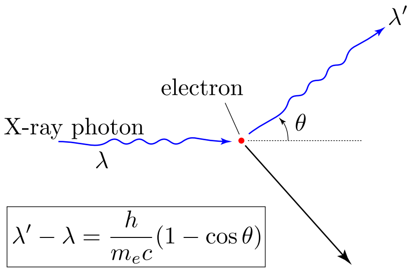


Fig. 3.37 Compton scattering of a X-ray photon off of an electron.

(a) What is the typical wavelength of X-ray photons? What is their momentum and what is their energy?

(b) How does the momentum and energy of the X-ray photon compare to the momentum and energy of an electron in the ground state of a Bohr hydrogen atom?

(c) Argue though the photon interacts both with the electron and the proton of the hydrogen atom, the proton does not see much action, the photon primarily interacts with the electron.

(d) Argue why the electron in the ground state of a hydrogen atom can be effectively considered "free" and "at rest" compared to the X-ray photon. Note that "ordinary" photons in the visible spectrum have very low momenta, but X-ray photons are not so ordinary!

(e) Upon scattering from the electrons (Fig. 3.37), the shift in the wavelength of the X-ray photon is given by the famous result $\lambda' - \lambda = \frac{h}{m_e c}(1-\cos\theta)$ where θ is the angle by which the photon is scattered. This expression is one of the first in which the mass of electron m_e, Planck's constant h, and the speed of light c appeared together! Show that this formula for the wavelength shift is consistent if we enforce the conservation of net momentum, and the conservation of net energy before and after the collision.

(f) Note that due to the wave-particle duality, the change in the wavelength is related to $\frac{h}{m_e c}$, which has the Planck's constant, the mass of the electron m_e, and the speed of light c. Since this quantity has the dimensions of Planck's constant by momentum, it must have units of wavelength by de Broglie relation. The quantity

$$\lambda_c = \frac{h}{m_e c} \tag{3.48}$$

is called the Compton wavelength, and shows up in electron scattering events at high energies. Calculate its magnitude, and make a plot of the X-ray wavelength shift $\lambda' - \lambda$ as a function of the scattering angle θ.

(3.7) Heisenberg's uncertainty principle

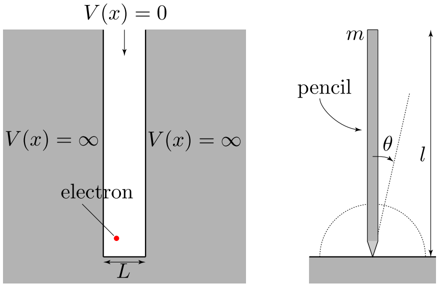


Fig. 3.38 A quantum particle in a box, and a quantum version of trying to balance a pencil on its tip.

The uncertainty principle is a powerful tool for obtaining estimates and orders of magnitudes of physical quantities in the quantum world. There are three uncertainty relations: the linear coordinate – linear momentum uncertainty relation $\Delta x \Delta p_x \geq \frac{\hbar}{2}$, the angular coordinate – angular momentum uncertainty relation $\Delta\theta\Delta L \geq \frac{\hbar}{2}$, and the time-energy uncertainty relation $\Delta t \Delta E \geq \frac{\hbar}{2}$. While the first two have similarities, the third is rather distinct. In this problem we will apply the uncertainty relation to a few examples to illustrate their usage, and in the process gain some insight into what they really mean.

(a) In Fig. 3.38, an electron of mass m_e is shown trapped in a 1D box (fancier name = quantum well) of length L. Invoking *only* the uncertainty principle $\Delta x \Delta p_x$ and the classical kinetic energy of the electron, argue why the energy of the electron cannot be zero, i.e., it cannot sit at the bottom of the well. Your quantitative argument should give you lower bounds on the electron momentum $\Delta p_x \geq \frac{\hbar}{2L}$, and the electron kinetic energy: $E_{kin} \geq \frac{\hbar^2}{8 m_e L^2}$. Estimate the two lower bounds for the electron if the quantum well size is $L \sim 1$ nm.

(b) Fig. 3.38 shows the interesting case of trying to balance a pencil of mass m and length l on its tip in the presence of gravity, which causes an acceleration g. First, using classical notions of angular momentum, show that if the initial angular velocity is $\omega_0 = \frac{d\theta}{dt}|_{t=0}$, then for small angles, the angle at time t is $\theta(t) \approx \omega_0 \tau_0 \sinh(\frac{t}{\tau_0})$, where $\tau_0 = \sqrt{\frac{2l}{3g}}$. The moment of inertia of a cylinder rotating around its end is $I = \frac{1}{3} m l^2$. Then, show that the uncertainty relation $\Delta\theta\Delta L \geq \frac{\hbar}{2}$ implies an initial angular velocity of at least $\omega_0 \geq \frac{3\hbar}{2\pi m l^2}$. Then estimate the time it will take due to the uncertainty relation for the pencil to fall over: you should get $t_{fall} \sim \tau_0 \ln(\frac{\pi^2 m l^2}{3\hbar\tau_0})$. Estimate the time for a pencil of length $l \sim 10$ cm and mass $m \sim 10$ g – you should get t_{fall} to be a few seconds!

(c) In Section 3.2, we found how Bohr, by obtaining the allowed electron energies, explained the spectral lines of hydrogen. A characteristic length scale that emerged from his arguments was the Bohr radius a_B. Use the uncertainty relation to estimate the energy of an electron that is trapped in a 3D cubic box, each side side of which is a Bohr radius long. You should get an energy of the order $\sim \frac{\hbar^2}{m_e a_B^2}$ with some numerical factors of the order unity. Note that this is the quantum confinement energy, and is of the same order of magnitude as the Rydberg or Bohr energy. However, this energy is positive, whereas the Bohr energy is negative. Why?

(d) If an electron is sitting in the $n = 2$ level of hydrogen, it will ultimately transition to the state $n = 1$ by emitting a single photon. Using the energy-time uncertainty, show that the **lower limit** of the time it will take for this transition to occur is $\tau \geq \frac{8 m_e a_B^2}{3\hbar}$, where m_e is the electron mass and a_B the Bohr radius. Verify that this time is in the 10s of femtoseconds. Now experimentally the transition from upper to lower levels and emission of a photon takes much longer, about 1 ns. Why?

(e) Now the *spontaneous* transition of an electron from a higher to lower energy eigenvalue should be troubling, because we learned in the solution of the time-dependent Schrödinger equation that energy eigenvalues are *stationary states*. If I put an electron in energy level E_2, it should stay there forever! Why does the electron make the transition to the lower level, i.e. from $E_2 \to E_1$? Does this process violate the stationary state principle? This is a deeper conceptual question, in answering it you will learn more about what stationary states actually mean.

(3.8) **Pauli's exclusion and the Aufbau principle**
Electron eigenvalues of the elements of the periodic table follow a filling of orbitals dictated by the Pauli exclusion principle, and the qualitative Aufbau priniciple. The order is shown in Fig. 3.39 schematically. The notation $|nx\rangle$ indicates the orbital in shell n and orbital type x. For example, $|3s\rangle$ indicates the s orbital in the third shell. The number of electrons that can occupy the $|1s\rangle$ orbital are 2, with wavefunctions $\psi_{1s,\uparrow}(r)$ and $\psi_{1s,\downarrow}(r)$. Each orbital has a spin degeneracy of $g_s = 2$. There are three $|np\rangle$ orbitals, giving the $|np\rangle$ orbitals a total electron filling capacity of 6. The $|nd\rangle$ orbitals can accommodate 10 electrons, and the $|nf\rangle$ orbitals can hold 14 electrons.

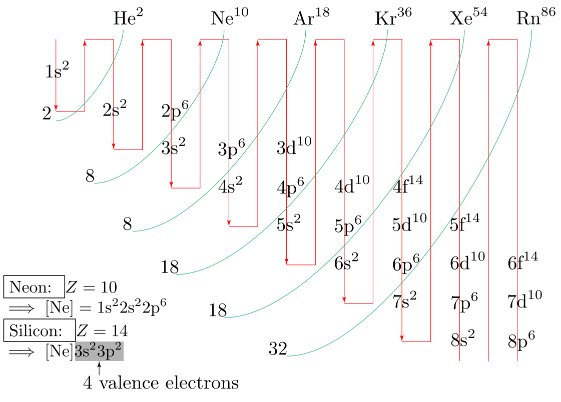


Fig. 3.39 The Aufbau principle and the Pauli exclusion for electron eigenvalues of the elements.

(a) Each electron state is labeled by four quantum numbers: (n, l, m, s). n is the shell number or the principal quantum number. The orbital quantum number can take values $l = 0, ..., n-1$. The azimuthal quantum number can take values $m = -l, ..., 0, ..., +l$, and the spin quantum number can be $s = \pm\frac{1}{2}$. Show that as a result, shell n can hold $2n^2$ electrons.

(b) The inert gases in the periodic table have filled outermost or valence orbitals. Show that this is the case for helium which has just two electrons. Except helium, all other inert gases have 8 electrons in their outermost shell. Identify the inert gases, their electron shell structure, and explain using the Pauli exclusion principle why these elements are inert.

(c) The examples of the electron shell structure of neon and silicon are shown. Note that the silicon atom has electrons in the $|3s\rangle$ and $|3p\rangle$ orbitals, and the p orbital is incomplete. Write the corresponding electron shell structures of C, Ge, and Sn. These elements from group IV form semiconductor crystals by forming covalent bonds: argue why sharing of 4 outermost electrons makes each atom appear as close to inert it can. This is a good **rule of** 8: if the sum of valence electrons is 8, chemical bonding can create semiconductors.

(d) Write the electronic shell structures of group III elements B, Al, Ga, In, and the group V elements N, P, As, and Sb. By investigating the electron shell structures, discuss why group III-V compound semiconductors such as GaN, GaAs, or InSb can be formed, driven by the "rule of 8".

(e) Based on the above principles, read and describe why there are so many elemental metals, yet so few elemental ferromagnets (e..g. Fe, Co, and Ni).

(3.9) **Klein–Gordon and Dirac's insomnia**
By the mid 1920s, Einstein's relativity established that the total energy of an object, be it particle or wave, is related to its rest mass m_0 and its momentum p through the relation

$$E^2 = (m_0 c^2)^2 + (pc)^2, \tag{3.49}$$

where c is the speed of light. For objects that have mass and are at rest, the famous $E = m_0 c^2$ result indicates a large amount of stored energy. The Schrödinger equation

$$i\hbar \frac{\partial}{\partial t}\psi = \hat{H}\psi \implies i\hbar \frac{\partial}{\partial t}\psi_E = E\psi_E \tag{3.50}$$

has no speed of light c in it, so it is inconsistent with relativity. This version of the Schrödinger equation is called the non-relativistic wave equation. Klein and Gordon obtained a relativistic wave equation, which applies to certain objects (such as the Higgs boson), but *not* to electrons. In this problem, we will derive this equation, and will glimpse how it led Dirac to find the origin of electron **spin**.

(a) Show that operating the time-dependent energy operator $-i\hbar\frac{\partial}{\partial t}$ *twice* to the wavefunction ψ, and in particular to energy eigenfunctions ψ_E leads to:

$$(i\hbar\frac{\partial}{\partial t})^2\psi = \hat{H}^2\psi \implies -\hbar^2\frac{\partial^2}{\partial t^2}\psi_E = [(m_0c^2)^2 + (pc)^2]\psi_E \tag{3.51}$$

(b) By rearranging the above equation, obtain the Klein-Gordon relativistic wave equation:

$$[\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2 + \frac{1}{\lambda^2}]\psi_E = 0, \tag{3.52}$$

where $\lambda = \frac{\hbar}{m_0c}$ is related to our familiar friend the Compton length from Equation 3.48 by a factor of 2π. Note closely the structure of this wave equation: it looks very similar to the Maxwell wave equation for electric and magnetic fields, but has the quantum content in the λ.

(c) Bohr thought that Klein and Gordon had successfully made quantum mechanics consistent through relativity, but Dirac disagreed. In fact, Schrödinger himself in search of the wave equation had come upon the Klein–Gordon form, but had rejected it because it led to negative probabilities, i.e. it had solutions for which $\psi^2 < 0$. It was later recognized that the Klein–Gordon equation does not apply to electrons, but to certain classes of particles that have no spin (such as Higgs bosons). Read about where the Klein–Gordon equation is applicable, and what does it help explain.

(d) Paul Dirac fought hard to solve the inconsistency of the Klein–Gordon equation, as he was after a relativistic wave equation that applied to the electron. He was convinced that the correct relativistic wave equation should be closer to Schrödinger's non-relativistic form because the non-relativistic form was so successful in explaining the spectra of hydrogen atoms. However, argue why writing Schrödinger equation simply as

$$i\hbar\frac{\partial}{\partial t}\psi_E = \sqrt{(m_0c^2)^2 + (pc)^2}\psi_E \tag{3.53}$$

Dirac must have ran into the problem of having to take the *square root* of an *operator*.

1902
P A M
DIRAC OM
PHYSICIST
iγ·∂ψ=mψ
1984

Fig. 3.40 Memorial stone honoring Dirac and his equation in the Westminster Abbey.

(e) It is said that Dirac pondered about this problem for quite some time, and one late evening staring into his fireplace, realized that the wavefunction cannot be a scalar function, but a 4×1 *matrix*. This led to the Dirac equation, a relativistic wave equation for the electron (Fig. 3.40). The Dirac equation explained the origin of the electron spin, predicted the existence of antimatter, and solved the negative probability problem of the Klein-Gordon form. Read about Dirac's discovery, and write a short report on its genesis, what it explained, what it predicted, and why it is so beloved in all fields of physics.

(3.10) **Fermions, Bosons, "Any"ons?**
The wave-particle duality and the probabilistic interpretation force the wavefunction squared to remain unchanged for two indistinguishable particles in a particle swap, requiring $\psi(r_2, r_1) = \pm\psi(r_1, r_2)$. This requirement resulted in us concluding that all indistinguishable quantum particles must either be fermions with the $-$ sign upon a swap, or bosons with the $+$ sign upon the swap. A third possibility that still leaves $|\psi(r_2, r_1)|^2 = |\psi(r_1, r_2)|^2$ is when $\psi(r_2, r_1) = e^{i\theta}\psi(r_1, r_2)$, where θ is a phase.

(a) Show that swapping $(r_1 \leftrightarrow r_2)$ *twice* brings the 2-particle wavefunction $\psi(r_1, r_2)$ back to itself for *both* fermions and bosons.

(b) Show that if swapping $(r_1 \leftrightarrow r_2)$ twice does not bring the wavefunction back to itself, the wavefunction must be of the form $\psi(r_2, r_1) = e^{i\theta}\psi(r_1, r_2)$, where θ is a phase. For this to happen, θ can be *any* angle *except* $\theta = n\pi$ where $n = 0, \pm 1, ...$, giving them the name anyons.

(c) This phase has a topological interpretation. Refer to Fig. 3.41 and discuss why the order in which the swap operation is performed matters, and how the phase θ can capture this information. This procedure is called ”braiding”.

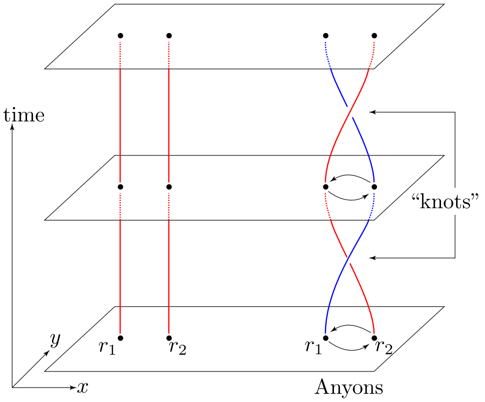


Fig. 3.41 Braiding properties of Anyons.

(d) For one class of anyons, the statistics is called ”Abelian” in which case the phase factor $e^{i\phi}$ commutes for a swap and its exact inverse. This case is believed to have been observed in semiconductors in a phenomenon called the fractional quantum Hall effect when electrons are confined in **two dimensions**. In this remarkable effect, the statistics of the particles seems to hover between fermionic and bosonic. Read about Abelian anyons and their manifestation in the FQHE, and write a short paragraph on what is strange about it. We will discuss some more about this effect in Chapter 25.

(e) Another class of anyons is called ”non-Abelian”, for whom the phase factor $e^{i\phi}$ fails to commute for a swap and its exact inverse. Explore non-Abelian anyons - these are elusive and though theoretically predicted, are yet to be observed in nature!

(3.11) **Planck's formula and the big bang**

Planck's formula for blackbody radiation has enabled us to peer into the past of our universe in the following way. In 1964, Arno Penzias and Robert Woodrow Wilson of the Bell laboratories built a high-sensitivity radio antenna for satellite communications. In some of their first measurements, they observed an excess $T = 4.2$ K antenna temperature background radiation that they could not explain. There were theoretical predictions that the big-bang must create a **cosmic microwave background** in the Planck's blackbody spectra, as a time-stamp of the origin of the universe, and it was realized that Penzias and Wilson had indeed measured this background. Plot the Planck's blackbody radiation formula at $T = 2$ K and $T = 4$ K and read about the historical development and implications of this discovery stemming from Planck's radiation formula.

More recently, it was also discovered by pointing more sensitive radio telescopes to various directions in space that the cosmic microwave background radiation pattern is not isotropic. The NASA cosmic background explorer (COBE) mission was able to image this anisotropy showing how Planck's radiation formula is accurate enough to detect minute variations in temperatures, and provides a tool to peer into the history of the universe since its birth in far more detail. Read and write a short essay about this more recent development.

Damned Lies and Statistics

4

4.1 Quantum statistics and entropy 59
4.2 The physics of equilibrium 62
4.3 Partition function for quantum systems 64
4.4 The Fermi–Dirac distribution 65
4.5 The Bose–Einstein distribution 66
4.6 Properties of the distribution functions 66
4.7 Quantum twist on thermodynamics 71
4.8 Meaning of equilibrium in semiconductor devices 72
4.9 Chapter summary section 76
Further reading 77
Exercises 77

In this chapter, we derive and discuss the Fermi–Dirac distribution function for fermions, and the Bose–Einstein distribution function for bosons. These functions provide us the statistical occupation number of quantum states for a system in thermodynamic equilibrium with a reservoir. The Fermi–Dirac distribution is central to finding the electron distribution over allowed energy or momentum values in various semiconductor devices. The Bose–Einstein distribution is central to finding the distribution of photons in the electromagnetic field, or phonons in semiconductor crystals. The two distributions together determine electron-phonon and electron-photon interactions. The importance of this chapter simply cannot be overemphasized! We discuss various properties of the distributions and limiting cases to gain familiarity. We then see how the equipartition of energy result of classical thermodynamics gets a facelift due to the quantum statistics. Then, we specifically map the concept of thermodynamic equilibrium to the fundamental semiconductor building blocks, such as the ohmic contact, Schottky contacts, the p-n junction, and a field-effect transistor (FET). So the goals of this chapter are to understand:

- Why is it necessary to update the Maxwell–Boltzmann distribution of particle energies?
- How does quantum mechanics combined with thermodynamics lead to the Fermi–Dirac and the Bose–Einstein distributions?
- What are the properties of the Fermi–Dirac and Bose–Einstein statistics that differ significantly from the classical Maxwell–Boltzmann statistics, and under what conditions are they the same?
- When and how should we apply the quantum statistics in semiconductor materials and devices?

Fig. 4.1 Ludwig Boltzmann, the father of statistical mechanics and kinetic theory of gases. He discovered the formula for entropy or disorder as a mathematical concept: this formula $S = k_b \log\Omega$ is inscribed on his gravestone. The concept of entropy permeates all branches of sciences, including electronic communication systems and stock markets. Boltzmann ended his life tragically by hanging himself in 1906.

4.1 Quantum statistics and entropy

The functional form in which a large number of particles distribute in energy, when subjected to the laws of quantum mechanics, is the Fermi–Dirac function for fermions, and the Bose–Einstein function for bosons. These distribution functions can be derived by the principle of **maximizing the entropy** of a large number of such particles. This principle is rooted in the "three" laws of thermodynamics:

Zeroth Law: If systems A and B are separately in thermodynamic equilibrium with system C, then A is in thermodynamic equilibrium with B. The zeroth law **defines the temperature** T as the quantity that is the same for systems in thermodynamic equilibrium.

First Law: The total energy dU supplied to a system divides into a part $dQ = TdS$ that changes the internal heat of the system, and the rest $dW = Fdx + \mu dN$ is the total mechanical work Fdx + chemical work μdN done on the system. Here T is the temperature, $S = k_b \ln \Omega$ is the Boltzmann entropy[1] (Fig. 4.1), F is the mechanical force, and dx a spatial displacement, μ is the chemical potential, and dN is the change in the number of particles. Thus, $dU = dQ + dW = TdS + (Fdx + \mu dN)$ is a **statement of energy conservation**.

Second Law: The second law formulated by Clausius (Fig. 4.3) states that spontaneous processes result in an **increase in entropy**, and to decrease entropy, work must be done. This is derived in Chapter 21. Complete conversion of heat energy to useful work is impossible as the sole result of any physical process. For example, there is no process whose sole result is the flow of energy from a cold to a hot body. For any closed cyclic process, $\oint \frac{dQ}{T} \geq 0$, the entropy cannot decrease.

Third Law: $S(T \to 0 \text{ K}) = 0$, or entropy goes to zero as $T \to 0$ K.

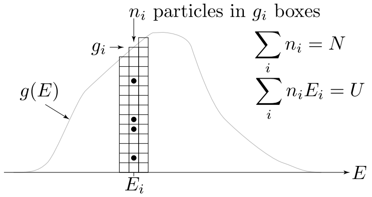


Fig. 4.2 The Fermi–Dirac and Bose–Einstein distributions are obtained by maximizing the Boltzmann entropy of distributing $N = \sum_i n_i$ particles of total energy $U = \sum_i n_i E_i$ in the energy states that are allowed by quantum mechanics. At energy E_i, there are g_i energy eigenvalues, which is given by the density of states $g(E)$. The distribution functions are independent of $g(E)$, and are given by $f(E_i) = \frac{n_i}{g_i}$.

Let us apply the maximization of entropy to derive the Fermi–Dirac, Bose–Einstein, and the Maxwell–Boltzmann distribution functions. Suppose there are g_i discrete energy eigenvalues allowed by quantum mechanics for energy E_i, defined by a density of states $g(E)$ as shown in Fig. 4.2. In these energy eigenvalues, we want to put N particles, where $N \sim 10^{23}$ is a large number. There are many possible ways of putting N particles in the allowed energy eigenvalues. If we can identify how many ways Ω there are, $S = k_b \ln \Omega$ is the Boltzmann entropy. Then, we can identify the particular distribution that results in the largest entropy, subject to the total number $N = \sum_i n_i$, and the total energy $U = \sum_i n_i E_i$. Though there are several distributions $\Omega_i = \{n_i\}$ that satisfy the particle number and energy requirements, the distribution that maximizes the entropy $\{n_i\}_0$ has an *overwhelmingly* large number of configurations (Exercise 4.1). Thus, we can neglect all distributions that have non-maximal entropy, and retain only the distribution $\{n_i\}_0$ that maximizes the entropy. This result is exact for $N \to \infty$.

Fermions: Because of the Pauli exclusion principle, each energy eigenvalue box in Fig. 4.2 can hold either 0, or a maximum of 1 fermionic particle. The number of distinct ways to put n_i identical non-interacting fermions in g_i boxes is $\frac{g_i!}{n_i!(g_i - n_i)!}$, which makes the total number of ways of putting all $N = \sum_i n_i$ particles to be[2]

[1] The Boltzmann constant k_b was actually introduced specifically by Max Planck. Boltzmann's version was in the form of the ideal gas constant R from the relation $PV = RT$ of gas theory, where $R = Nk_b$, and N the number of atoms. Planck's blackbody radiation formula discussed in Chapter 3 shone the spotlight on k_b. Ω is the number of distinct configurations of the system as described later.

[2] Here, the Stirling approximation $\ln[N!] \approx N \ln[N] - N$ for large N is used extensively (see Exercise 4.13).

$$\Omega = \prod_i \frac{g_i!}{n_i!(g_i - n_i)!}$$

$$S = -k_b \ln \Omega = -k_b \sum_i \left[\ln(g_i!) - \ln(n_i!) - \ln((g_i - n_i)!) \right]$$

$$\approx -k_b \sum_i g_i \Big[\frac{n_i}{g_i} \ln(\frac{n_i}{g_i}) + (1 - \frac{n_i}{g_i}) \ln(1 - \frac{n_i}{g_i}) \Big]. \tag{4.1}$$

This entropy is maximized by subtracting from it the requirements of fixed particle number $\alpha(\sum_i n_i - N)$ and fixed total energy $\gamma(\sum_i n_i E_i - U)$, where α and γ are Lagrange multiplier constants, and finding the maxima w.r.t. n_i. Writing the ratio $\frac{n_i}{g_i} = f_i$ as the occupation fraction of energy state E_i, we get:

Fig. 4.3 Rudolf Clausius formulated the second law of thermodynamics, and in the process introduced the concept (and the word) entropy into physics.

$$\frac{\partial}{\partial n_i} \Big[S - \alpha(\sum_i n_i - N) - \gamma(\sum_i n_i E_i - U) \Big] = 0$$

$$-k_b \Big[\ln(f_i) - \ln(1 - f_i) \Big] - (\alpha + \gamma E_i) = 0$$

$$\implies f_i = \frac{1}{e^{\frac{\alpha + \gamma E_i}{k_b}} + 1} \underbrace{\implies}_{\alpha = -\frac{\mu}{T}, \gamma = \frac{1}{T}} \boxed{f_i = f(E_i) = \frac{1}{e^{\frac{E_i - \mu}{k_b T}} + 1},} \tag{4.2}$$

where the boxed form is the **Fermi–Dirac distribution** for fermionic particles. The constants α and γ are identified from the particle number and energy constraints, and μ is a **chemical potential**[3]. The entropy of the distribution of the Fermi particles is then:

$$S = -k_b \sum_i g_i \Big[f_i \ln f_i + (1 - f_i) \ln(1 - f_i) \Big]$$

$$S_n = -k_b \sum_k \Big[f_k \ln f_k + (1 - f_k) \ln(1 - f_k) \Big], \tag{4.3}$$

where the first form is in terms of the density of states in energy $g_i = g(E_i)$, whereas the second form is a sum over the fermion wavevectors k for a band of index n and energy eigenvalues $E_n(k)$. In anticipation of future chapters where allowed electron energies in a crystal arrange into bands $E_n(k)$, the entropy sum is simply[4] over the k-states, S_n is the entropy of band of index n, and $f_k = 1/(\exp[\frac{E_n(k) - \mu}{k_b T}] + 1)$.

Bosons: Since there is no exclusion principle for bosons, there is no restriction on how many particles can be put in any eigenvalue, or box in Fig. 4.2. The number of ways of putting n_i bosons into g_i boxes is then $\frac{(g_i + n_i - 1)!}{n_i!(g_i - 1)!}$, as may be seen in Fig. 4.4. Following a similar entropy maximization process for Fermions, we obtain the **Bose–Einstein dis-**

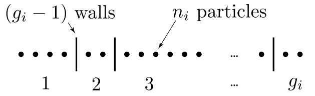


Fig. 4.4 The number of ways or arranging n_i identical particles in g_i boxes with no restriction on particles/box reduces to the number of ways of inserting $g_i - 1$ walls between n_i sites. The total number of (particles + walls) is $g_i + n_i - 1$.

---

[3] The chemical potential μ is a measure of the particle number. It is *defined* by this derivation. The symbol μ is the same as mobility, but confusion should be avoided based on the context!

[4] Note that the net electronic entropy of a completely filled band with $f_k = 1$, or a completely empty band $f_k = 0$ are both zero. This is expected, since they are both perfectly ordered states with only one possible way of occurring, i.e., $\Omega = 1$.

tribution and the corresponding entropy:

$$\boxed{f_i = f(E_i) = \frac{1}{e^{\frac{E_i-\mu}{k_b T}} - 1}}$$

$$S = -k_b \sum_i g_i \Big[f_i \ln(f_i) - (1+f_i)\ln(1+f_i) \Big]$$

$$S = -k_b \sum_k \Big[f_k \ln(f_k) - (1+f_k)\ln(1+f_k) \Big]. \tag{4.4}$$

Boltzmann Distribution: Both the Fermi–Dirac and Bose–Einstein distributions reduce to the Boltzmann distribution at high temperatures and/or with very few particles, when the distributions become non-degenerate, as discussed later in this chapter[5]. If we remove the requirement of indistinguishability, the restrictions placed by quantum mechanics on wavefunctions for Fermions or Bosons vanish. If we further assume $n_i << g_i$, then the number of ways of arranging the particles in energy state E_i is $g_i^{n_i}/n_i!$. Following entropy maximization, we obtain the **Boltzmann distribution** and its entropy[6]

$$\boxed{f_i = f(E_i) = e^{\frac{\mu - E_i}{k_b T}}}$$

$$S = -k_b \sum_i g_i f_i \ln f_i$$

$$S = -k_b \sum_k f_k \ln f_k. \tag{4.5}$$

Looking back at how exactly the three boxed distributions were obtained, we recognize that the quantity that was maximized in Equation 4.2 was not the entropy S alone, but $S - \frac{(U-\mu N)}{T}$. Dividing the relation $W = (U - \mu N) - TS$ from the first law of thermodynamics by the temperature T, we recognize that the distribution functions are exactly those configurations of particles that **minimize the free energy**, or the "work" W at a fixed T. This is achieved by maximizing the entropy and minimizing the internal energy under the constraint of fixed particle number and total energy: the three equilibrium distributions are indeed the "laziest" possible arrangements of the particles! These distributions are dragged into states of non-equilibrium by external electric fields, which will be the topic of discussion of Chapter 21.

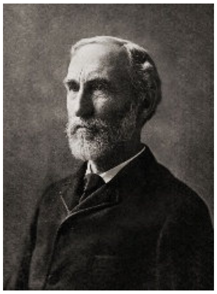


Fig. 4.5 Josiah Willard Gibbs coined the word "statistical mechanics", and with Maxwell and Boltzmann, gave it its modern rigorous mathematical basis. He developed the modern form of vector calculus. Gibb's quiet lifestyle was in stark contrast to the eternal impact of his scientific ideas and work.

4.2 The physics of equilibrium

The distribution functions may also be derived by appealing to properties of statistical thermodynamics, via the grand partition function. The fundamental assumption on which all of statistical mechanics rests is simple to state: **All accessible quantum states of a system are equally probable**. We begin exploring its quantitative consequences by drawing upon a fundamental result from quantum statistical mechanics. The most well-known result of statistical thermodynamics is

[5] Note that for bosons the chemical potential μ is non-zero for particles whose number is constant (such as atoms and molecules), and $\mu = 0$ for particles whose number is not constant (such as photons or phonons). Exercise 4.4 discusses this aspect, and the difference of μ for fermions and bosons further.

[6] Note that both Fermi–Dirac and Bose–Einstein distributions asymptotically approach the Boltzmann distribution for $(E_i - \mu)/k_b T >> +1$, as illustrated later in this chapter. The formulation of entropy where we sum over the k-index of allowed states instead of the energies E_i shows a direct correspondence with Shannon's formulation of *information* entropy, $S = -\sum_i p_i \ln(p_i)$, where p_i is the probability of an event.

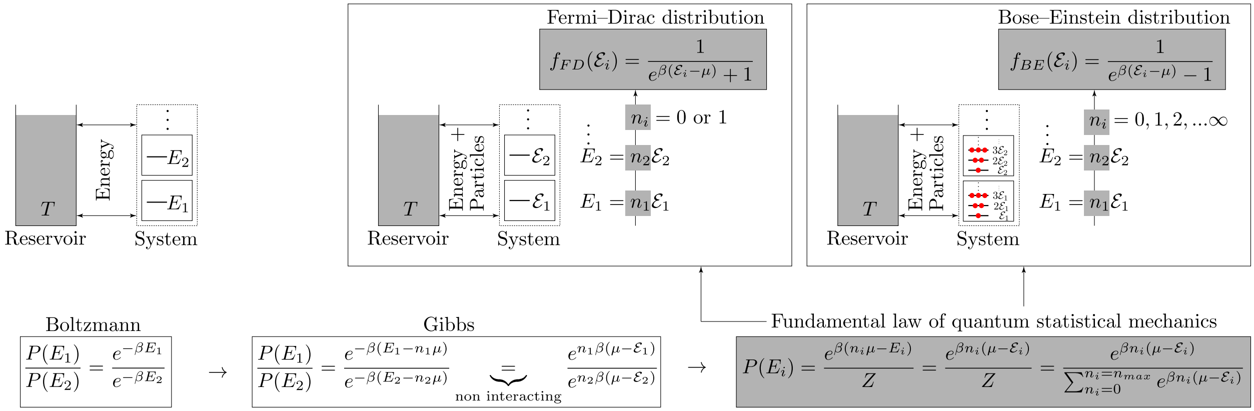

Reservoir
System
T
Energy
E_2
E_1
Fermi–Dirac distribution
$f_{FD}(\mathcal{E}_i) = \frac{1}{e^{\beta(\mathcal{E}_i-\mu)}+1}$
Energy + Particles
$n_i = 0$ or 1
$E_2 = n_2\mathcal{E}_2$
$E_1 = n_1\mathcal{E}_1$
Bose–Einstein distribution
$f_{BE}(\mathcal{E}_i) = \frac{1}{e^{\beta(\mathcal{E}_i-\mu)}-1}$
$n_i = 0, 1, 2, ...\infty$
$3\mathcal{E}_2$ $2\mathcal{E}_2$ $\mathcal{E}_2$
$3\mathcal{E}_1$ $2\mathcal{E}_1$ $\mathcal{E}_1$
Boltzmann
$\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta E_1}}{e^{-\beta E_2}}$
Gibbs
$\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta(E_1-n_1\mu)}}{e^{-\beta(E_2-n_2\mu)}} \underbrace{=}_{\text{non interacting}} \frac{e^{n_1\beta(\mu-\mathcal{E}_1)}}{e^{n_2\beta(\mu-\mathcal{E}_2)}}$
Fundamental law of quantum statistical mechanics
$P(E_i) = \frac{e^{\beta(n_i\mu-E_i)}}{Z} = \frac{e^{\beta n_i(\mu-\mathcal{E}_i)}}{Z} = \frac{e^{\beta n_i(\mu-\mathcal{E}_i)}}{\sum_{n_i=0}^{n_i=n_{max}} e^{\beta n_i(\mu-\mathcal{E}_i)}}$

Fig. 4.6 Illustration of the processes of thermodynamic equilibrium for the Boltzmann distribution, and the Gibbs partition function.

the Boltzmann distribution. The result states the following: consider a *system* that is in *thermal equilibrium* with a *reservoir* at temperature T. Each of the terms in italics have very specific meanings, which will be described shortly. Let E_1 and E_2 denote two energy states of the system. The Boltzmann result asserts that the probabilities of finding the system in these energies is related by

$$\boxed{\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta E_1}}{e^{-\beta E_2}},} \tag{4.6}$$

where $\beta = \frac{1}{k_b T}$, and k_b is the Boltzmann constant. Fig. 4.6 illustrates the meanings of the terms in italics. The *reservoir* is a large source of particles and energy, characterized by a temperature T. It goes by the name *reservoir* because it can either take in, or give out, any energy without changing its temperature T. As opposed to the reservoir, the *system* is much smaller, and can be found in energy states E_1, E_2, E_3, ... The statement that the system is in *thermal equilibrium* with the reservoir means that it can exchange energy with the reservoir, but *not particles*. Each energy state E_i is considered to be *individually* in thermal equilibrium with the reservoir. Only under this condition is the Boltzmann result in Equation 4.6 applicable. Since the temperature T is a measure of the energy which is being exchanged, the reservoir and the system share the same temperature upon reaching thermal equilibrium.

Now if we let the system exchange energy *and particles* with the reservoir, as indicated in Fig. 4.6, the Boltzmann relation needs to be generalized. A measure of the particle number is the chemical potential μ, which must also appear in addition to the temperature T in relations characterizing thermodynamic equilibrium between the system and the reservoir. This famous generalization was done by Gibbs

(Fig. 4.5), who gave the modified relation

$$\boxed{\frac{P(E_1)}{P(E_2)} = \frac{e^{-\beta(E_1 - n_1\mu)}}{e^{-\beta(E_2 - n_2\mu)}} \underbrace{=}_{\text{non-interacting}} \frac{e^{n_1\beta(\mu-\mathcal{E}_1)}}{e^{n_2\beta(\mu-\mathcal{E}_2)}},} \tag{4.7}$$

where μ is a common chemical potential of the reservoir+system, and n_i is the number of particles in the single-particle energy state $\mathcal{E}_i$. We are going to call a single particle energy eigenstate an *orbital*, drawing from the language of chemistry. Only if the particles considered are *non-interacting*, then the energy of the state is $E_i = n_i\mathcal{E}_i$ if there are n_i particles in orbital $|i\rangle$ of eigenvalue $\mathcal{E}_i$. If these conditions are met, then one defines a Gibbs-sum, popularly known as the **grand partition function:**

$$Z = \sum_{\text{states}} \sum_{\text{n}} e^{\beta(n\mu - E_n)}. \tag{4.8}$$

The sum runs over all states of the system, and all number of particles allowed in each single-particle state. Note carefully what this means. For example, consider the situation when orbital $|3\rangle$ is in equilibrium with the reservoir. Since it is not interacting with the other orbitals (which are also separately in equilibrium with the reservoir), the partition function for the "system" consisting of a variable number of particles in $|3\rangle$ is then $Z = \sum_{n_3=0}^{n_3=n_{max}} e^{\beta n_3(\mu-\mathcal{E}_3)}$. The "system" here is the various occupation states of orbital $|3\rangle$.

When both energy *and* particle exchange is allowed between the system and the reservoir, the fundamental law of equilibrium statistical mechanics may be stated as the following. Under thermodynamic equilibrium with a reservoir at temperature T, the absolute probability that the system will be found in the state $E_i = n_i\mathcal{E}_i$ with n_i particles in orbital $|i\rangle$ is

$$\boxed{P(E_i) = \frac{e^{\beta(n_i\mu - E_i)}}{Z} = \frac{e^{\beta n_i(\mu-\mathcal{E}_i)}}{Z} = \frac{e^{\beta n_i(\mu-\mathcal{E}_i)}}{\sum_{n_i=0}^{n_i=n_{max}} e^{\beta n_i(\mu-\mathcal{E}_i)}}.} \tag{4.9}$$

4.3 Partition function for quantum systems

For sake of completeness and for future use, we generalize the above result for quantum systems. This section may be skipped in an initial reading. We recognize that the allowed orbital energies of any quantum system $\mathcal{E}_i$ are the eigenvalues of the Schrödinger equation. This may be written formally using the single-particle Hamiltonian $\hat{H}_0$ via $\hat{H}_0|i\rangle = \mathcal{E}_i|i\rangle$, the non-interacting many-particle Hamiltonian $\hat{H} = \sum \hat{H}_0$ gives $\hat{H}|n_1, n_2, ...n_i, ...\rangle = (\sum_i n_i\mathcal{E}_i)|n_1, n_2, ...n_i, ...\rangle$, and the number n_i of particles in the eigenstate (or orbital) $|i\rangle$ is $\hat{N}_i|n_1, n_2, ...n_i, ...\rangle = n_i|n_1, n_2, ...n_i, ...\rangle$, where $\hat{N}_i$ is occupation number operator for eigenstate $|i\rangle$, and $\hat{N} = \sum_i \hat{N}_i$. Then, the expectation value of any operator $\langle\hat{O}\rangle$ at thermodynamic equilibrium is

$$\boxed{\langle \hat{O} \rangle = \frac{\mathrm{Tr}\left[\hat{O} e^{\beta(\mu \hat{N} - \hat{H})}\right]}{\mathrm{Tr}\left[e^{\beta(\mu \hat{N} - \hat{H})}\right]},} \tag{4.10}$$

where Tr[...] stands for the trace of a matrix as discussed in Chapter 6. Note that the Hamiltonian matrix and the number operator are exponentiated. The trace gives the sum of the diagonal elements, making Equation 4.10 equivalent to 4.9 in the diagonal representation. But since the trace is invariant between representations, Equation 4.10 also holds for non-diagonal conditions. Feynman[7] calls the fundamental results in Equation 4.9 (and 4.10) the ”summit of statistical mechanics, and the entire subject either a slide-down from the summit, or a climb up to this result”. We have not covered the climb-up, but since we will apply the result, let us slide down by applying it to derive the Fermi–Dirac and the Bose–Einstein distribution functions. We will not need the services of Equation 4.10 here, Equation 4.9 will suffice for our current needs.

Fig. 4.7 Richard Feynman is one of the most well-known physicists because of his colorful lifestyle. He made significant contributions to several areas of physics, and developed the path integral approach to quantum mechanics, as a method distinct from Schrödinger and Heisenberg approaches. His contributions to quantum electrodynamics and the eponymous Feynman diagrams won him the 1965 physics Nobel Prize.

4.4 The Fermi–Dirac distribution

As we have discussed in Chapter 3, the number of fermionic particles that can occupy an energy eigenstate $\mathcal{E}_i$ are $n_i = 0$ or 1 and *nothing else* because of the **Pauli exclusion principle**. Therefore, the partition function for the state of the system corresponding to energy $\mathcal{E}_i$ in thermodynamic equilibrium (in the Gibbs sense) with a reservoir of temperature T and chemical potential μ is simply

$$Z = \sum_{n_i=0}^{n_i=1} e^{\beta n_i(\mu - \mathcal{E}_i)} = e^0 + e^{\beta(\mu - \mathcal{E}_i)} = 1 + e^{\beta(\mu - \mathcal{E}_i)}, \tag{4.11}$$

and the probability that the system is in a state that has n_i particles in orbital $|i\rangle$ is simply $P(E_i) = e^{\beta(n_i\mu - E_i)}/Z$, where $E_i = n_i\mathcal{E}_i$ is the total energy of the orbital. Note that we are assuming that the particles do not interact with each other. Then, the thermal average number of

---

[7] *Statistical Mechanics*, by R. P. Feynman. About his Nobel Prize in 1965, Feynman recounts: ”I was in the Cornell cafeteria and some guy, fooling around, throws a plate in the air. As the plate went up in the air I saw it wobble, and I noticed the red medallion of Cornell on the plate going around. It was pretty obvious to me that the medallion went around faster than the wobbling. I had nothing to do, so I start figuring out the motion of the rotating plate. I discovered that when the angle is very slight, the medallion rotates twice as fast as the wobble rate – two to one. It came out of a complicated equation! I went on to work out equations for wobbles. Then I thought about how the electron orbits start to move in relativity. Then there’s the Dirac equation in electrodynamics. And then quantum electrodynamics. And before I knew it the whole business that I got the Nobel Prize for came from that piddling around with the wobbling plate.” A replica of the Cornell plate is now part of an exhibit marking the centennial of the Nobel Prize. Feynman’s Messenger lecture series at Cornell are highly recommended, and can be viewed here: https://bit.ly/3xNUTH9.

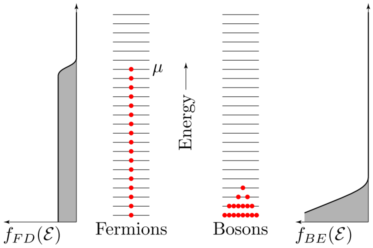


Fig. 4.8 The Pauli exclusion principle for fermions forces particles to fill orbitals of increasing energy when the particle number is increased. The chemical potential μ is therefore a direct measure of the number of fermions. For bosons, there is no exclusion principle and one can put as many particles at any energy.

particles $\langle n_i \rangle$ in orbital $|i\rangle$ is given by $f(\mathcal{E}_i) = \langle n_i \rangle = \sum_i n_i P(E_i)$:

$$\langle n_i \rangle = f(\mathcal{E}_i) = \frac{0 \cdot e^0 + 1 \cdot e^{\beta(1\cdot\mu - 1\cdot\mathcal{E}_i)}}{1 + e^{\beta(\mu - \mathcal{E}_i)}} \implies \boxed{f_{FD}(\mathcal{E}_i) = \frac{1}{e^{\beta(\mathcal{E}_i - \mu)} + 1}.} \tag{4.12}$$

The boxed form is the same Fermi–Dirac distribution as in Equation 4.2, obtained from a different point of view. It varies between 0 and 1, and is equal to $\frac{1}{2}$ when $\mathcal{E}_i = \mu$. More properties are discussed later.

4.5 The Bose–Einstein distribution

Unlike the Pauli exclusion principle for Fermions, there is no restriction on the number of Bosonic particles that can occupy an orbital $|i\rangle$. This means $n_i = 0, 1, ..., \infty$ are allowed. Then, the partition function is

$$Z = \sum_{n_i=0}^{\infty} e^{\beta n_i(\mu - \mathcal{E}_i)} = \sum_{n_i=0}^{\infty} [e^{\beta(\mu - \mathcal{E}_i)}]^{n_i} = \frac{1}{1 - e^{\beta(\mu - \mathcal{E}_i)}}, \tag{4.13}$$

where the infinite sum is a geometric series $1 + u + u^2 + ... = \frac{1}{1-u}$, valid for $u = e^{\beta(\mu - \mathcal{E}_i)} < 1$, or equivalently $\mu \leq \mathcal{E}_i$. The thermal average number of bosonic particles in orbital $|i\rangle$ is then

$$\langle n_i \rangle = f(\mathcal{E}_i) = \frac{0 \cdot u^0 + 1 \cdot u^1 + 2 \cdot u^2 + ...}{(1-u)^{-1}} \implies \boxed{f_{BE}(\mathcal{E}_i) = \frac{1}{e^{\beta(\mathcal{E}_i - \mu)} - 1}.} \tag{4.14}$$

The boxed form is the Bose–Einstein distribution, same as Equation 4.4. Here we have identified $u\frac{d}{du}(\frac{1}{1-u}) = \frac{u}{(1-u)^2} = u + 2u^2 + 3u^3 + ...$ to be the sum that appears in the numerator, whereupon $\langle n_i \rangle = \frac{1}{u^{-1}-1}$. Note that for $\beta(\mathcal{E}_i - \mu) >> 1$, the Bose–Einstein distribution $f_{BE}(E_i) \to 0$. However, for $\beta(\mathcal{E}_i - \mu) << 1$, $f_{BE}(E_i) \approx \frac{1}{\beta(\mathcal{E}_i - \mu)}$ can increase without bound: surprisingly this has a physical manifestation in the *condensation* of all particles to the lowest energy orbitals[8].

4.6 Properties of the distribution functions

The key ideas and results in arriving at the distribution functions are summarized in Figs. 4.6 and 4.8. In Fig. 4.9, all three distribution functions are plotted together for comparison. Note their differences in the quantum regime, and similarities in the classical regime.

For the rest of this section we discuss properties of the Fermi–Dirac function. Defining the function

$$f_0(x) = \frac{1}{1 + e^{\beta x}}, \tag{4.15}$$

the argument $x = E - \mu$ gives us the Fermi–Dirac distribution

$$f_{FD}(E) = f_0(E - \mu) = \frac{1}{1 + e^{\beta(E - \mu)}}. \tag{4.16}$$

---

[8] This phenomenon is related to the Bose–Einstein condensation of bose-particles that results in phenomena such as superfluidity.

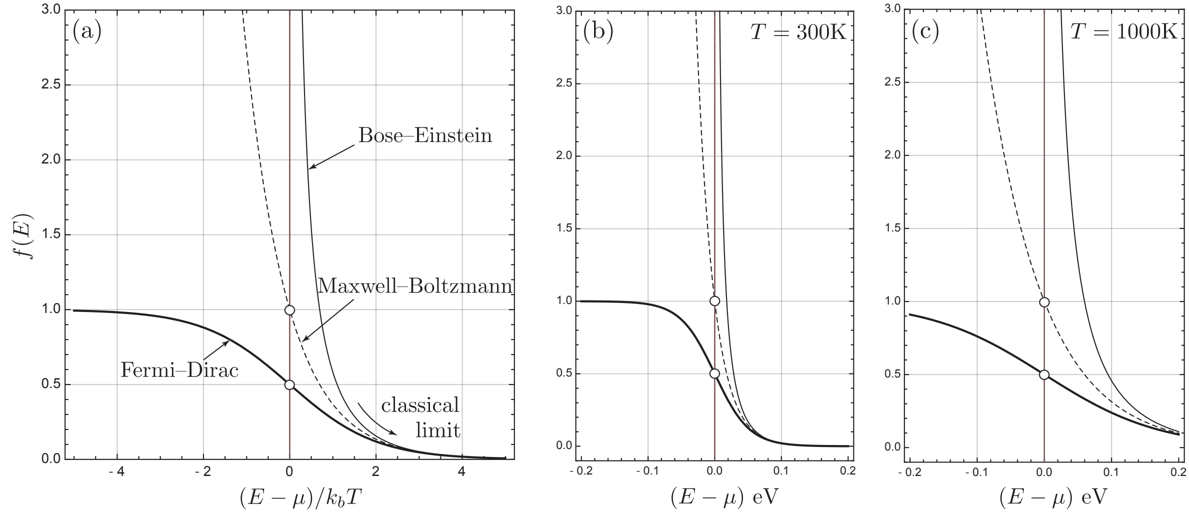


Fig. 4.9 (a) Illustrates all three distribution functions, and (b, c) illustrate the effect of temperature on them. The Fermi–Dirac occupation stays below 1, whereas there is no such limit for the Bose–Einstein occupation. Both Fermi–Dirac and Bose–Einstein distributions merge to the Maxwell–Boltzmann function at high energies $E >> \mu$, marking the quantum to classical transition.

The distribution may be thought of a function of the energy E, or of the chemical potential μ. We use the compact notation $f_0 = f_0(E-\mu) = f_{FD}(E)$. The partial derivative with respect to energy is

$$\frac{\partial f_0}{\partial E} = -\frac{\partial f_0}{\partial \mu} = -\beta \cdot \frac{e^{\beta(E-\mu)}}{(1+e^{\beta(E-\mu)})^2} = -\beta \cdot f_0[1-f_0], \tag{4.17}$$

which can be rearranged to the form

$$\boxed{-\frac{\partial f_0}{\partial E} = +\frac{\partial f_0}{\partial \mu} = \frac{\beta}{4\cosh^2(\frac{\beta(E-\mu)}{2})}.} \tag{4.18}$$

Because $\cosh^2(x) \geq 1$, the derivative of the Fermi–Dirac distribution reaches its maximum value of $\frac{\beta}{4} = \frac{1}{4k_bT}$ at $E = \mu$. The integral of this function $\int_{-\infty}^{+\infty} du\beta/(4\cosh^2[\frac{1}{2}\beta u]) = 1$ indicates that in the limit of very low temperatures $\frac{1}{k_bT} = \beta \to \infty$, the derivative function should approach a Dirac-delta function in the energy argument[9], i.e.,

$$\boxed{\lim_{T\to 0}[-\frac{\partial f_0}{\partial E}] = \lim_{T\to 0}[+\frac{\partial f_0}{\partial \mu}] = \delta(E-\mu).} \tag{4.19}$$

This feature is illustrated in Fig. 4.10.

Now considering $f(u) = 1/(1+e^u)$ and $f(v) = 1/(1+e^v)$, we get the identity

$$\boxed{f(u) - f(v) = \underbrace{[f(u)+f(v)-2f(u)f(v)]}_{\geq 0} \times \tanh(\frac{v-u}{2}).} \tag{4.20}$$

---

[9] Because of the Pauli Exclusion principle, the electrons in a solid that contribute to electrical conductivity are from a small window of energies where the Fermi derivative function peaks: the electronic conductivity σ is proportional to $-\frac{\partial f_0}{\partial E}$. This is because states of lower energy are completely occupied, and states with high energies have no electrons. This is crudely the same effect as when we blow air on a bowl of water, the surface responds, not the interior. This is derived quantitatively in later chapters.

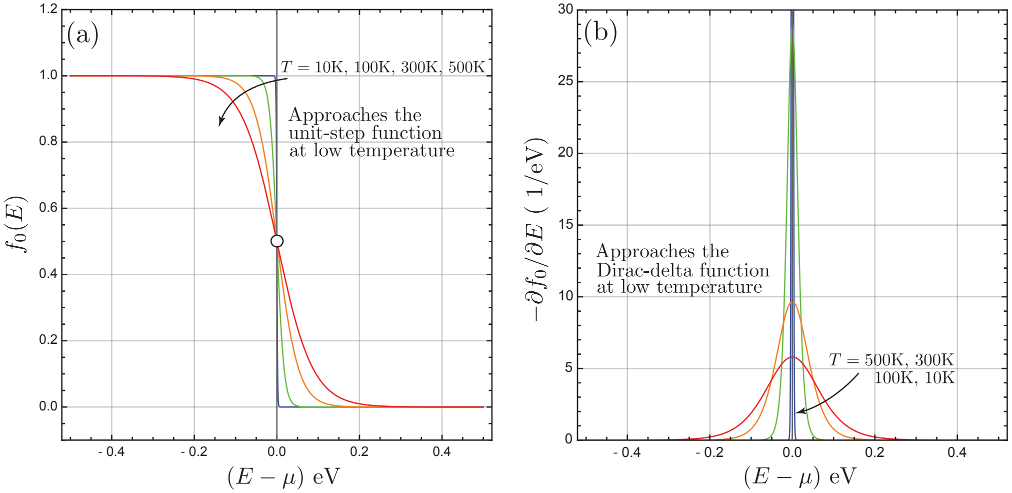


Fig. 4.10 Illustration of (a) the temperature dependence of the Fermi–Dirac distribution, and (a) its derivative. While the Fermi–Dirac distribution approaches a unit-step at low temperatures, its derivative takes becomes sharply peaked, approaching the Dirac-delta function in the limit $T \to 0$K.

Since for fermions $f(u), f(v) \leq 1$, the term in the square brackets is always positive. So the sign of the Fermi difference function is determined by the tanh(...) term. The Fermi difference function will appear repeatedly when we study the optical and electronic transport properties of semiconductors and electronic and photonic devices[10].

The integral of the Fermi–Dirac function is

$$\boxed{\int_0^\infty dE f_0(E-\mu) = \int_0^\infty \frac{dE}{1+e^{\beta(E-\mu)}} = \frac{1}{\beta}\ln(1+e^{\beta\mu}),} \tag{4.21}$$

which leads to the very useful Fermi *difference* integral

$$\boxed{\int_0^\infty dE[f_0(E-\mu_1) - f_0(E-\mu_2)] = (\mu_1-\mu_2) + \frac{1}{\beta}\ln[\frac{1+e^{-\beta\mu_1}}{1+e^{-\beta\mu_2}}].} \tag{4.22}$$

If $\mu_1, \mu_2 >> k_b T$, the second term on the right side is zero[11], and we obtain

$$\boxed{\int_0^\infty dE[f_0(\mu_1) - f_0(\mu_2)] \approx (\mu_1-\mu_2).} \tag{4.23}$$

That this relation is an identity is evident at $T \to 0$, or $\beta \to \infty$. Features of the Fermi difference function are illustrated in Fig. 4.11. The integral at low temperatures is just the area under the dashed difference curve, which is rectangular and has a energy width of $\mu_2 - \mu_1$.

It will prove very useful later to define higher moment integrals of the Fermi–Dirac function of the form

$$F_j(\eta) = \frac{1}{\Gamma(j+1)}\int_0^\infty du \frac{u^j}{1+e^{u-\eta}}. \tag{4.24}$$

[10] For example, in Chapter 29 the switching of sign of the Fermi difference function will be critical to the creation of population inversion in a laser.

[11] The Fermi difference function will appear later in the problem of electrical current flow. Electrons in a metal or a semiconductor are subjected to two Fermi levels when connected to a battery. The electrons in the Fermi difference function window are those responsible for current flow.

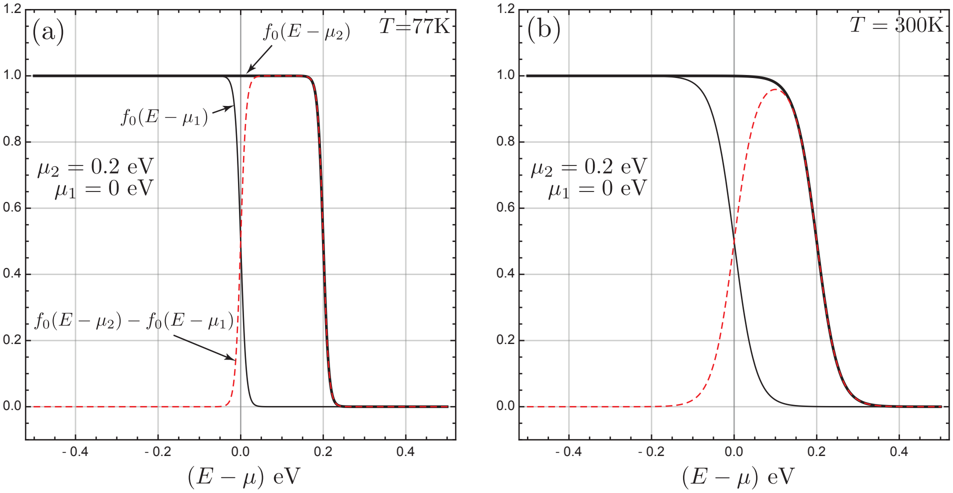


Fig. 4.11 Illustration of the temperature dependence of the Fermi difference distribution at (a) $T = 77$K and (b) at $T = 300$K. The difference function peaks in an energy window $\mu_1 \leq E \leq \mu_1$. It becomes increasingly rectangular as the temperature drops.

The Fermi–Dirac integral is rendered dimensionless by scaling the chemical potential $\eta = \beta\mu$, and the energy $u = \beta E$ by the thermal energy $k_b T = \frac{1}{\beta}$. Since we are integrating over u, the Fermi–Dirac integral $F_j(\eta)$ is a function of the chemical potential μ. The denominator is a normalizing Gamma function $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$ with the property $\Gamma(n+1) = n\Gamma(n)$, which means if n is an integer, $\Gamma(n) = (n-1)!$. A useful value of the Gamma function for a non-integer argument is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. For $\eta << -1$, the exponential in the denominator is much larger than unity. An excellent approximation of the Fermi–Dirac integral then is $F_j(\eta) \approx e^\eta$, irrespective of the value of j. In the other extreme, when $\eta >> +1$, an excellent approximation to leading order is $F_j(\eta) \approx \frac{\eta^{j+1}}{\Gamma(j+2)}$. Due to the high importance of Fermi–Dirac integrals in semiconductor devices, we collect the results[12]:

$$\boxed{F_j(\eta) = \frac{1}{\Gamma(j+1)}\int_0^\infty du \frac{u^j}{1+e^{u-\eta}},} \boxed{F_j(\eta) \underbrace{\approx}_{\eta<<-1} e^\eta,} \boxed{F_j(\eta) \underbrace{\approx}_{\eta>>1} \frac{\eta^{j+1}}{\Gamma(j+2)}.} \tag{4.25}$$

From Equation 4.21, we have an exact analytical result for the Fermi–Dirac integral for $j = 0$: it is $F_0(\eta) = \ln(1 + e^\eta)$. The validity of the approximations in Equation 4.25 are easily verified for this special case. No exact analytical expressions for other orders ($j \neq 0$) exist. The approximations in Equation 4.25 then assume increased importance for analytical evaluation of various physical quantities such as the mobile carrier densities in semiconductor bands, transport phenomena, and optical properties. The order j depends on the dimensionality of the problem. Fig. 4.12 illustrates the cases of the Fermi–Dirac integrals and their approximations for the cases of $j = 0$ and $j = \frac{1}{2}$.

[12] The Fermi–Dirac integrals in Fig. 4.12 are central to understanding the operation of electronic switches, or transistors. The current flowing in a transistor is proportional to a Fermi–Dirac integral in the y-axis, while the voltage controlling the current is proportional to the x-axis. As a result, we will see in later chapters that vast, *orders of magnitude* changes in current can be obtained by small changes in the voltage. The high-current is the on-state, and the low-current will be the off-state of the transistor switch.

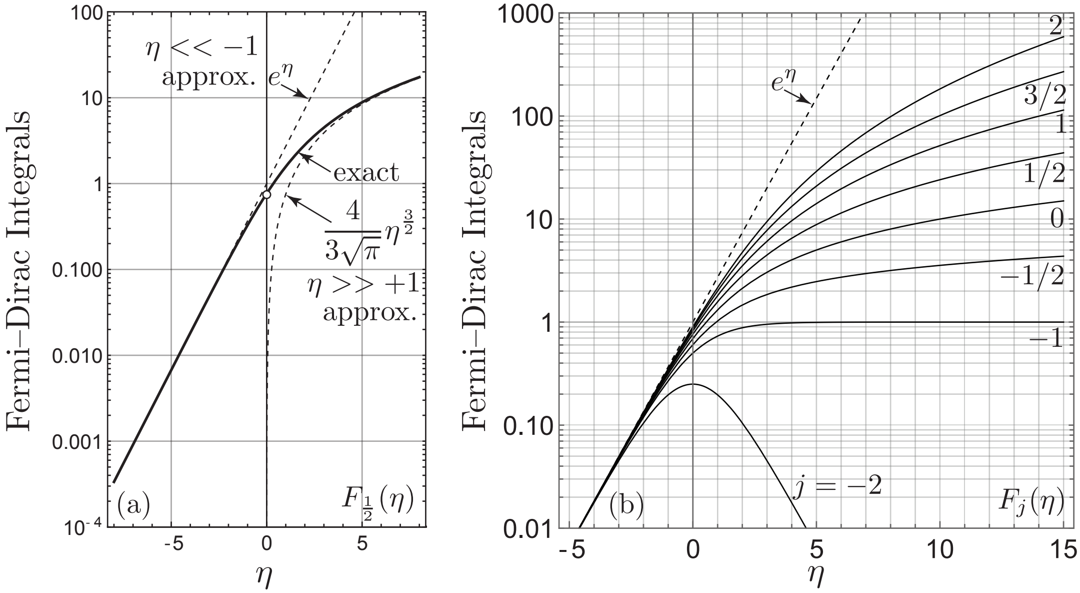


Fig. 4.12 Fermi–Dirac integrals and their non-degenerate ($\eta << -1$) and degenerate ($\eta >> 1$) approximations, illustrating Equation 4.25. (a) The integral $F_{\frac{1}{2}}(\eta)$ and its non-degenerate and degenerate approximation. (b) The integral $F_j(\eta)$ for several j, and can be used for looking up the numerical values of the integrals.

Table 4.1 The Gamma function $\Gamma(n)$ and the Riemann zeta function $\zeta(n)$ for a few arguments.

n	$\Gamma(n)$	$\zeta(n)$
-2	∞	0
$-\frac{3}{2}$	$\frac{4\sqrt{\pi}}{3}$	-0.025
-1	∞	$-\frac{1}{12}$
$-\frac{1}{2}$	$-2\sqrt{\pi}$	-0.208
0	∞	$-\frac{1}{2}$
$+\frac{1}{2}$	$\sqrt{\pi}$	-1.460
$+1$	1	∞
$+\frac{3}{2}$	$\frac{\sqrt{\pi}}{2}$	$+2.612$
$+2$	1	$\frac{\pi^2}{6}$
$+\frac{5}{2}$	$\frac{3\sqrt{\pi}}{4}$	$+1.341$
$+3$	2	$+1.202$
$+\frac{7}{2}$	$\frac{15\sqrt{\pi}}{8}$	$+1.127$
$+4$	6	$\frac{\pi^4}{90}$
$+\frac{9}{2}$	$\frac{105\sqrt{\pi}}{16}$	$+1.055$
$+5$	24	$+1.037$

We will rarely need higher order terms of the Fermi–Dirac integral other than the approximations in Equation 4.25. For the rare cases where they are necessary, it has been shown that the Fermi Dirac integral is given by the complete expansion

$$\boxed{F_j(\eta) \underbrace{=}_{\eta>0,j>-1} \sum_{n=0}^{\infty} \frac{2t_{2n}}{\Gamma(j+2-2n)}\eta^{j+1-2n} + \cos(\pi j)\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n^{j+1}}e^{-n\cdot\eta},} \tag{4.26}$$

where $t_0 = 1/2$, $t_n = (1-2^{1-n})\zeta(n)$, and $\zeta(n)$ is the Riemann zeta function. Some representative values of the Riemann zeta function and the Gamma function are listed in Table 4.1. We also note that for $j = 0$, $F_0(\eta) = \ln(1+e^{\eta})$ is an exact result. Since the Fermi–Dirac integrals follow the identity

$$\frac{\partial}{\partial\eta}F_j(\eta) = F_{j-1}(\eta), \tag{4.27}$$

and $F_0(\eta) = \ln(1+e^{\eta})$, the Fermi–Dirac integral is analytical for $j = 0, -1, -2, \ldots$. For example, $F_{-1}(\eta) = \frac{\partial}{\partial\eta}F_0(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$, and so on. In the degenerate case, $\eta > 0$, the second summation term on the right in Equation 4.26 can be neglected because it drops as powers of $e^{-\eta}$, and the first sum can be approximated to be

$$F_j(\eta) \underbrace{\approx}_{\eta >> 0, j > -1} \frac{\eta^{j+1}}{\Gamma(j+2)} + \frac{\pi^2}{6\Gamma(j)} \eta^{j-1} + ... \tag{4.28}$$

Table 4.2 lists the degenerate approximations of the Fermi–Dirac integral up to the second term for a few orders. Note that $\Gamma(0)$ is not defined, but is not necessary since $F_0(\eta) = \ln(1+e^{\eta})$ is exact.

Table 4.2 Approximations for the Fermi–Dirac integral in the degenerate limit $\eta >> 0$ for a few orders. The second, and subsequent terms of the expansions are much smaller than the first when $\eta >> 0$.

j	$F_j(\eta)$
$-\frac{1}{2}$	$\approx \frac{2}{\sqrt{\pi}} \eta^{\frac{1}{2}} - \frac{\pi^{\frac{3}{2}}}{12} \eta^{-\frac{3}{2}} + ...$
0	$= \ln(1+e^{\eta}) \approx \eta$
$+\frac{1}{2}$	$\approx \frac{4}{3\sqrt{\pi}} \eta^{\frac{3}{2}} + \frac{\pi^{\frac{3}{2}}}{6} \eta^{-\frac{1}{2}} + ...$
$+1$	$\approx \frac{1}{2}\eta^2 + \frac{\pi^2}{6} + ...$
$+\frac{3}{2}$	$\approx \frac{8}{15\sqrt{\pi}} \eta^{\frac{5}{2}} + \frac{\pi^{\frac{3}{2}}}{3} \eta^{\frac{1}{2}} + ...$
$+2$	$\approx \frac{1}{6}\eta^3 + \frac{\pi^2}{6}\eta + ...$
$+\frac{5}{2}$	$\approx \frac{16}{105\sqrt{\pi}} \eta^{\frac{7}{2}} + \frac{2\pi^{\frac{3}{2}}}{9} \eta^{\frac{3}{2}} + ...$
$+3$	$\approx \frac{1}{12}\eta^4 + \frac{\pi^2}{12}\eta^2 + ...$

4.7 Quantum twist on thermodynamics

A hallmark result of classical mechanics and classical thermodynamics is the principle of the **equipartition of energy**. This result asserts that if a classical system allows several modes (or degrees of freedom), the average energy allotted to each mode is exactly $\frac{1}{2}k_bT$. The result is evident if we consider a large number of particles of the same mass m that are free to move in one dimension (=1 degree of freedom)[13]. The kinetic energy of a particle moving with velocity v_x is then $E = \frac{1}{2}mv_x^2$. The Maxwell-Boltmann distribution function asserts that for classical particles in equilibrium with a reservoir at temperature T, to minimize the system energy, the probability of finding a particle with a kinetic energy $E = \frac{1}{2}mv_x^2$ is $f(E) = \exp\left[-E/(k_bT)\right]$. This is a Gaussian distribution in velocity that attaches the highest probability to the slowest moving particles. The average energy is

$$\langle E \rangle_{1d} = \frac{\int_{-\infty}^{+\infty} dv_x \cdot (\frac{1}{2}mv_x^2) \cdot e^{-\frac{\frac{1}{2}mv_x^2}{k_bT}}}{\int_{-\infty}^{+\infty} dv_x \cdot e^{-\frac{\frac{1}{2}mv_x^2}{k_bT}}} = \frac{1}{2}k_bT. \tag{4.29}$$

If the particles are free to move in three dimensions, they can have velocities $\mathbf{v} = (v_x, v_y, v_z)$, leading to an average energy

$$\langle E \rangle_{3d} = \frac{\int_{-\infty}^{+\infty} dv_x \int_{-\infty}^{+\infty} dv_y \int_{-\infty}^{+\infty} dv_z \cdot (\frac{1}{2}m(v_x^2+v_y^2+v_z^2)) \cdot e^{-\frac{\frac{1}{2}m(v_x^2+v_y^2+v_z^2)}{k_bT}}}{\int_{-\infty}^{+\infty} dv_x \int_{-\infty}^{+\infty} dv_y \int_{-\infty}^{+\infty} dv_z \cdot e^{-\frac{\frac{1}{2}m(v_x^2+v_y^2+v_z^2)}{k_bT}}}$$
$$\implies \langle E \rangle_{3d} = \frac{1}{2}k_bT + \frac{1}{2}k_bT + \frac{1}{2}k_bT = \frac{3}{2}k_bT. \tag{4.30}$$

Note that each degree of freedom gets $k_bT/2$ energy. This is the equipartition of energy of classical thermodynamics. Now we must fix this result of the equipartition of energy to account for the quantum statistics that are obeyed by fermions and bosons. The procedure is as simple as replacing the Maxwell–Boltzmann distribution with the correct statistics for quantum particles. For non-interaction fermions, we get for 1D the quantum result

[13] A note about the mass term m: for example for electrons in vacuum m is the free electron mass m_e, but for electrons in solids it is the effective mass $m^{\star}$ as discussed in future chapters. The results of this section are completely general, and are not limited to electrons. For example m can be the mass of N_2 molecules for Nitrogen gas in a box.

$$\langle E\rangle_{1d} = \frac{\int_{-\infty}^{+\infty} dv_x \cdot (\frac{1}{2}mv_x^2) \cdot \frac{1}{e^{\frac{\frac{1}{2}mv_x^2-\mu}{k_bT}}+1}}{\int_{-\infty}^{+\infty} dv_x \cdot \frac{1}{e^{\frac{\frac{1}{2}mv_x^2-\mu}{k_bT}}+1}} = \frac{1}{2}k_bT \cdot \frac{F_{1/2}(\eta)}{F_{-1/2}(\eta)}. \qquad (4.31)$$

To evaluate the integrals, we have used a change of variables $E = \frac{1}{2}mv_x^2$, and the definitions of the Fermi–Dirac integrals defined in Equation 4.25, and $\eta = \mu/(k_bT)$. The Gamma functions give the factor of 1/2 in front because $\frac{\Gamma(1+\frac{1}{2})}{\Gamma(1-\frac{1}{2})} = \frac{\sqrt{\pi}/2}{\sqrt{\pi}} = \frac{1}{2}$. In the non-degenerate limit $\eta << -1$, $F_j(\eta) \approx e^{\eta}$, which immediately gives us the classical equipartition of energy result $\langle E\rangle_{1d} = \frac{1}{2}k_bT$. But in the quantum or degenerate limit, $F_j(\eta) \approx \frac{\eta^{1+j}}{\Gamma(2+j)}$, and we get the average energy as $\langle E\rangle_{1d} \approx \frac{k_bT}{2} \cdot \frac{\mu}{k_bT} \cdot \frac{\Gamma(\frac{3}{2})}{\Gamma(\frac{5}{2})} \approx \frac{1}{3}\mu$. This is the $T \to 0$ K limit of a fermi gas of particles. We will encounter this result again in Chapter 5.

For non-interacting Fermion particles moving in 3D, similarly, we get the quantum result

$$\langle E\rangle_{3d} = \frac{3}{2}k_bT \cdot \frac{F_{3/2}(\eta)}{F_{1/2}(\eta)}. \qquad (4.32)$$

The classical limit when $\eta << -1$ is $\langle E\rangle_{3d} = \frac{3}{2}k_bT$, but the quantum limit is $\langle E\rangle_{3d} = \frac{3}{5}\mu$. This is shown in Fig. 4.13. Exercise 4.2 asks you to generalize this quantum result for fermions with kinetic energy $\frac{1}{2}mv^2$ moving in d-dimensions where $d = 1, 2, 3$ to get the **quantum equipartition result for fermions**:

$$\boxed{\langle E\rangle_d = \frac{d}{2}k_bT \cdot \frac{F_{\frac{d}{2}}(\eta)}{F_{\frac{d-2}{2}}(\eta)}.} \qquad (4.33)$$

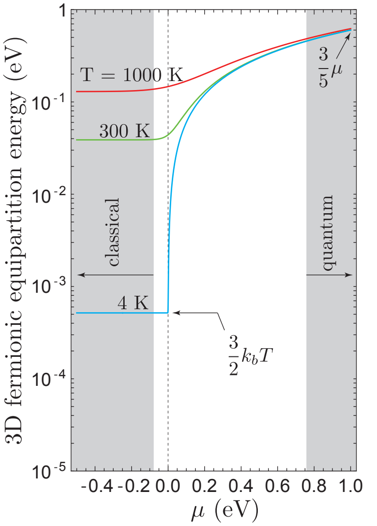


Fig. 4.13 The quantum version of the equipartition of energy result for 3D. For very high temperatures and for a low number of particles, collectively called the non-degenerate conditions, the classical energy $\frac{3}{2}k_bT$ is obtained. But for very large number of particles or at low temperatures, collectively called degenerate conditions, the quantum result is $\frac{3}{5}\mu$ for the average energy.

4.8 Meaning of equilibrium in semiconductor devices

You may skip this section in an initial reading. If you are familiar with semiconductors, you will find the discussion easygoing and useful. If you are not familiar with semiconductors, please come back to this section after reading about bandgaps and energy-band diagrams in later chapters.

Let us now consider the role of thermodynamics in a few semiconductor devices to develop a deeper understanding of the meaning of equilibrium in them. The first and simplest example is a 1D semiconductor (for example a carbon nanotube or a thin nanowire), which has *ohmic contacts* to two metal electrodes. The allowed energy eigenvalues in the semiconductor channel are those in the valence and conduction

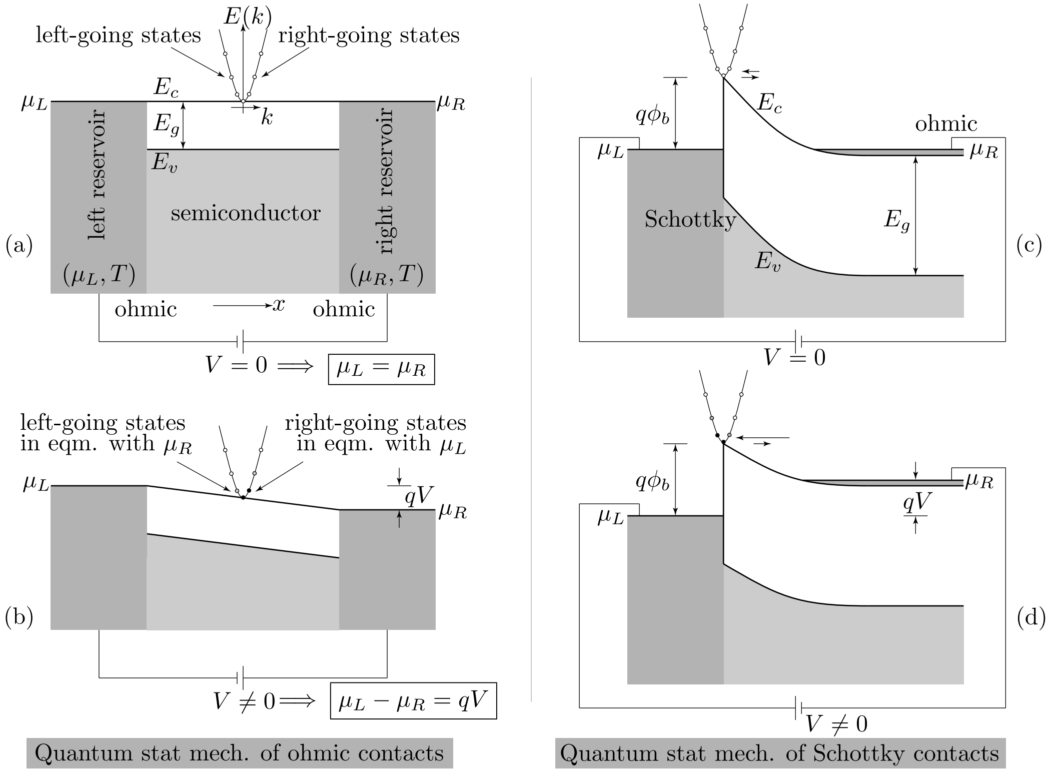

left-going states
$E(k)$
right-going states
E_c
μ_L
μ_R
k
E_g
E_v
left reservoir
semiconductor
right reservoir
(a)
(μ_L, T)
(μ_R, T)
ohmic
x
ohmic
$V = 0 \implies \boxed{\mu_L = \mu_R}$
left-going states in eqm. with μ_R
right-going states in eqm. with μ_L
μ_L
qV
μ_R
(b)
$V \neq 0 \implies \boxed{\mu_L - \mu_R = qV}$
Quantum stat mech. of ohmic contacts
$q\phi_b$
E_c
ohmic
μ_L
μ_R
Schottky
E_g
E_v
(c)
$V = 0$
$q\phi_b$
μ_R
μ_L
qV
(d)
$V \neq 0$
Quantum stat mech. of Schottky contacts

Fig. 4.14 Illustration of the concept of equilibrium for ohmic and Schottky contacts between metals and semiconductors. The left side shows an ohmic contact between electrons in the conduction band of a semiconductor with a source contact on the left and a drain contact on the right for (a) $V = 0$ and (b) $V \neq 0$. The corresponding cases for a Schottky contact are shown in (c) and (d).

bands with band edge energies E_v, E_c, separated by a bandgap E_g, as indicated in Fig. 4.14 (a). Consider the 1D semiconductor to be doped n-type, with mobile electrons in the conduction band, and no mobile carriers in the valence band. Then the true meaning of an *ohmic contact* is the following: the electrons in the conduction band of the semiconductor are in thermodynamic equilibrium with the electrons in the metal contacts, in the Gibbs sense. The conduction band states (or orbitals) in the semiconductor can freely exchange particles (electrons) and energy with the states or orbitals in the contacts, which is the reservoir. Connect this concept of Gibbs equilibrium in Fig. 4.14 (a) with the picture used earlier in Fig. 4.6. Note here we have *two reservoirs*. The particles in the left contact (reservoir) are in equilibrium with each other, and those in the right contact are in equilibrium with each other. When no external voltage is applied across them, the contacts are also in thermodynamic equilibrium with each other.

Inside the semiconductor connecting the contacts, there are particles that are moving to the right, and those moving to the left. Let us consider the situation where the left- and right-going carriers *do not mix*, i.e., there is no scattering of carriers. This is referred to as the *ballis-*

tic case, and is approximately realized for very short semiconductor lengths. Consider the electrons moving to the right in the semiconductor. These electrons can only enter the semiconductor from the left contact. Then the electrons moving to the right are in thermodynamic equilibrium with the *left contact*. Similarly, carriers moving to the left in the semiconductor are in equilibrium with the *right contact*. *Being in thermodynamic equilibrium in the Gibbs sense means the right-moving electron states share the same chemical potential μ and temperature T as the electrons in the left contact metal*. This is an extremely important consequence of thermodynamic equilibrium. Similarly the carriers in the semiconductor moving to the left could only have entered from the right contact, which keeps them in equilibrium with that contact: they share the same μ and the same T as the right contact. As long as the chemical potentials of the contacts are the same, the net current flow due the left and right moving carriers in the semiconductor exactly cancel, because they share the same μ.

When a voltage is applied between the contacts, the chemical potential of one contact is $\mu_L - \mu_R = qV$ larger than the other as shown in Fig. 4.14 (b). This in turn breaks the delicate balance of left-and right-moving carriers inside the semiconductor. The imbalance of the left and right moving carriers as indicated in Fig. 4.14 (b) thus is the driver of an electric current flow through the semiconductor, completing the circuit. We will use this picture to calculate the current through a ballistic semiconductor channel in Chapters 5 and 17, and show that the conductance is quantized. Then we will use this idea in Chapter 19 in the three-terminal ballistic transistor switch.

If the chemical potential of the metal lines up with energies in the bandgap of the semiconductor, a Schottky contact results, as indicated in Fig. 4.14 (c). The figure shows again a semiconductor in contact with two metals: the left contact is now Schottky, and the right contact is ohmic to the conduction band electrons in the semiconductor. Going back to our discussion of equilibrium in Section 4.2, we realize that the left moving electrons in the semiconductor are in thermodynamic equilibrium with the right contact. But the right moving electrons in the semiconductor are *not* in thermodynamic equilibrium with the left contact in the Gibbs sense, because there is a barrier between them that prevents free particle exchange. When a voltage is applied across the two metal contacts, the stronger imbalance of equilibrium between the left- and right-going carriers in the semiconductor cause a high asymmetry in the current flow as a function of the voltage. For the "forward" bias condition shown in Fig. 4.14 (d), the left moving carriers in the semiconductor that make it over the barrier to the metal are in equilibrium with the right contact. Since their chemical potential changes linearly with voltage, their concentration increases exponentially with voltage, causing a characteristic exponential turn-on of the diode. We will discuss this quantitatively in Chapter 18.

Fig. 4.15 shows the energy band diagrams of a semiconductor p-n junction. Note the reservoirs are metals, but clearly we have chosen

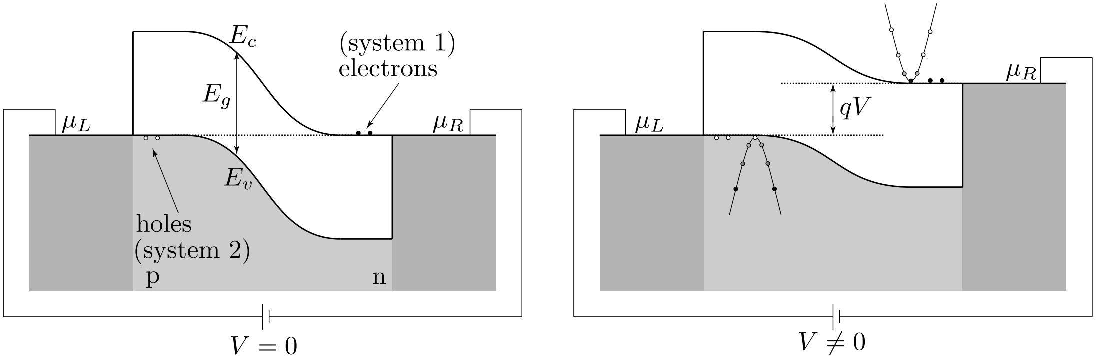


Fig. 4.15 Illustration of the concept of equilibrium for p-n junctions. Making ohmic contacts to the p-type semiconductor on the left and the n-type semiconductor on the right require metals with different work functions.

two *different* metals to form ohmic contacts individually to the p-side and the n-side of the semiconductor. An ohmic contact between a semiconductor and one metal electrode is possible for carriers in only one of the semiconductor bands, not both. This means with the proper choice of metals, we can form an ohmic contact to the conduction band of a n-type semiconductor for a n-type ohmic contact, and to the valence band of a p-type semiconductor for a p-type ohmic contact separately. So the holes in the valence band of the p-type semiconductor layer are in thermodynamic equilibrium with the p-ohmic metal (left), and the electrons in the conduction band of the n-type semiconductor layer are in thermodynamic equilibrium with the n-contact metal. Note now we also have two types of carriers – electrons in the conduction band, and holes in the valence band. When no voltage is applied, the holes in the p-side are in thermodynamic equilibrium with the electrons in the n-side – because they are in turn in equilibrium with their respective ohmic contact metal reservoirs. So they share a common chemical potential. However, when a voltage is applied, as indicated in Fig. 4.15, the equilibrium is broken; the chemical potentials of the conduction band electrons in the n-side and valence band holes in the p-type now differ by $\mu_R - \mu_L = qV$. This difference is responsible for current flow, as will be discussed in Chapter 18.

As a final example, consider a 3-terminal device, the field-effect transistor (FET) shown in Fig. 4.16. The carriers in the inversion channel form ohmic contacts to the source and the drain, but a Schottky-type contact through an additional insulating barrier layer to the gate metal reservoir. So the carriers in the semiconductor channel can be in thermal equilibrium with the carriers in the gate metal, but not in thermodynamic equilibrium with them in the Gibbs sense because the exchange of particles between the gate metal and the semiconductor channel is explicitly prohibited. The right-going carriers are again injected from the left contact, and the left-going carriers are injected from the right contact. But the carrier density in the semiconductor is

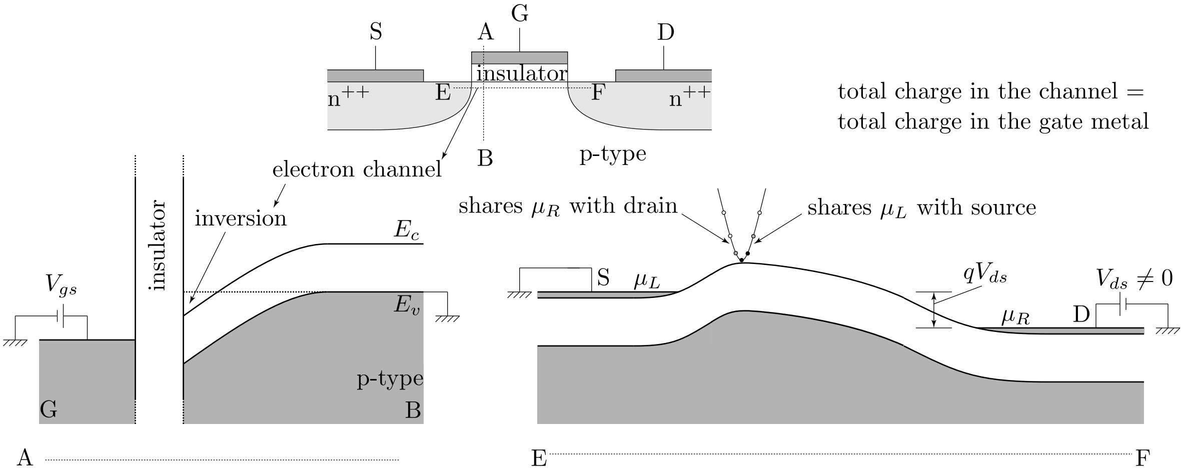

S
A
G
D
insulator
n++
E
F
n++
total charge in the channel =
total charge in the gate metal
electron channel
B
p-type
insulator
inversion
E_c
E_v
V_{gs}
p-type
G
B
shares μ_R with drain
shares μ_L with source
S
μ_L
qV_{ds}
$V_{ds} \neq 0$
μ_R
D
A
E
F

Fig. 4.16 Illustration of the concept of equilibrium for a 3-terminal MOSFET device. Of the electrons in the conduction band of the semiconductor inversion layer at the surface, the right-moving electrons are in equilibrium with the left contact, and the left-moving electrons are in equilibrium with the right contact. The gate voltage controls the number of electrons in the inversion channel.

controlled by the gate voltage capacitively. We will use this picture in Chapter 19 to discuss the ballistic FET in detail. The FET is the most commonly used semiconductor device today.

4.9 Chapter summary section

In this chapter, we learned:

- When many indistinguishable **fermions** such as electrons are allowed to distribute in energy, the occupation function is given by the Fermi–Dirac distribution $\boxed{f_{FD}(\mathcal{E}) = \frac{1}{e^{\beta(\mathcal{E}-\mu)}+1}.}$ Here $\beta = 1/k_bT$ indicates that the electrons are in Gibbs equilibrium with a reservoir of electrons at temperature T, and the number of fermions (or electrons) determines the chemical potential μ. Pauli exclusion principle clamps the maximum occupation of an orbital to 1.
- When many indistinguishable **bosons** such as photons or phonons are allowed to distribute in energy, the occupation function is given by the Bose–Einstein distribution $\boxed{f_{BE}(\mathcal{E}) = \frac{1}{e^{\beta(\mathcal{E}-\mu)}-1}.}$ The bosons of this system are in Gibbs equilibrium with a reservoir of bosons at temperature T.

- In semiconductors, electrons are the prototype fermions. The fundamental electronic and photonic properties will be predominantly determined by the Fermi–Dirac function. The Bose–Einstein distribution function will make its appearance when we discuss the interaction of photons with electrons, or of phonons with electrons. Photons and phonons are bosons. Excitons, which are electron-hole quasiparticles, behave as composite bosons.
- In semiconductor electronic and photonic devices, the occupation functions are controlled by applying voltages, currents, or shining light.

Further reading

Because the subject of statistical mechanics, thermodynamics, and entropy has deep philosophical implications, several good books have been written, and continue to be written on the topic. To learn the subject, **Statistical Physics** by Tony Guénault gives a gentle but no-nonsense introduction, and **Thermal Physics** by Kittel and Kroemer gives a solid introduction to the subject with several examples.

For the intermediate level, **Fundamentals of Statistical and Thermal Physics** by F. Reif, and **Statistical Physics: A Set of Lectures** by R. P. Feynman are highly recommended. For advanced level treatments, Volume 5 **Statistical Physics** of the Landau and Lifshitz Course of Theoretical Physics, and the two volume **Statistical Physics of Particles** and **Statistical Physics of Fields** by M. Kardar are the old and new classics.

Exercises

(4.1) **Meaning of entropy**
Fig. 4.17 indicates a lattice of N equivalent sites, on which we can put $n = 1, 2, ..., N$ identical particles, with the maximum of one particle per site.

(a) A visceral feel for entropy and its relation to disorder may be obtained by glancing at the three configurations, closing the book, and trying to recall which sites are occupied. This is simple *both* for the cases with very few particles (left), or a large number of particles (right). But it is very difficult for the middle case (unless you have a photographic memory!), where the number of particles is near half the sites. And this gets *very* difficult when N approaches very large numbers, of the order of $N = 10^{20}$. This is because there are far larger number of possible arrangements for $n = N/2$ than for $n = 1$ or $n = (N-1)$.

(b) Make a plot of the number of possible equivalent configurations Ω_n vs. the number of particles n on the lattice for N ranging from 10 to 1000. Note how fast the peak develops around $n = N/2$.

(c) Associate with the distributions the entropy via $S_n = k_b \ln \Omega_n$ and make a plot.

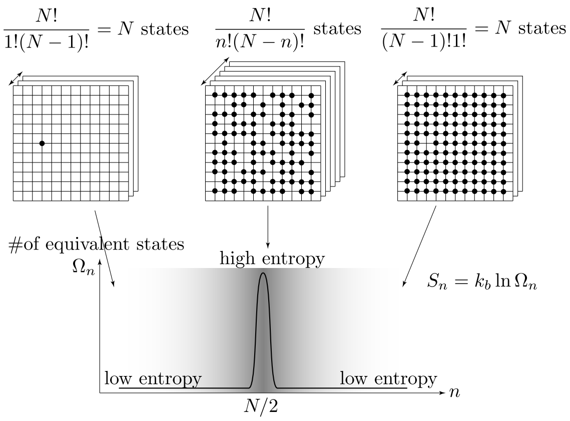


Fig. 4.17 Meaning of entropy.

(4.2) Quantum equipartition of energy

In Section 4.7, we learnt how the quantum statistics of fermions and bosons modify the classical equipartition of energy result. In this problem, we gain more acquaintance with this fundamental result that underlies much of semiconductor physics.

(a) Derive Equation 4.33 for fermions with kinetic energy $E = \frac{1}{2}mv^2$. Make plots of this energy for $d = 1, 2, 3$ similar to Fig. 4.13.

(b) Derive the corresponding result for bosons with kinetic energy $E = \frac{1}{2}mv^2$. Make plots of this energy for $d = 1, 2, 3$.

(c) Show that in the non-degenerate limit, both the Fermi–Dirac and Bose–Einstein distributions merge into the classical Maxwell–Boltzmann distribution function for all dimensions.

(4.3) Fermi–Dirac and Bose–Einstein distributions

(a) Phonons are bosons. A silicon crystal is sitting at room temperature $T = 300$ K, and $k_b T = 26$ meV. How many average number of phonons are there in a phonon mode of energy $\hbar\omega = 1$ meV?

(b) Electrons are fermions. I put $N = 100$ electrons in a harmonic oscillator potential whose quantum energy eigenvalues are $E_n = n\hbar\omega_0$ where $n = 1, 2, 3, ...$, and $\hbar\omega_0 = 0.2$ eV. At $T = 0$K, what is the Fermi energy E_F, and the effective velocity v_F of electrons at the Fermi energy?

(4.4) Chemical potential of fermions and bosons

The chemical potential μ of a system of particles is the change in the total system energy U due to the change in number of particles N, when all other parameters are held constant, i. e., $\mu = \frac{\partial U}{\partial N}$.

(a) For fermions, argue why at $T = 0$K the chemical potential μ must be equal to the highest energy occupied state. This energy is called the **Fermi energy** E_F, as is discussed in Chapter 5.

(b) At finite temperatures $T > 0$K, if the fermion number does not change, it will be seen in Chapter 5 that $E_F(T > 0\text{K}) \neq E_F(T = 0\text{K})$. In semiconductors, the number of fermions may split between different bands as electrons and holes. In such cases, it is convenient to define a **Fermi level** at finite temperatures at thermal equilibrium, which is different from $\mu = E_F(T = 0\text{K})$. Furthermore, to mimic systems out of thermal equilibrium such as a semiconductor device through which a current is flowing, the Fermi level will split into **quasi-Fermi levels**, each controlled by a separate contact or source of Fermions. Be aware of these variants of the concept of chemical potential of fermions, as they will appear in Chapter 5 and beyond.

(c) For bosons at $T = 0$K, argue why the concept of a chemical potential is not as straightforward as that for fermions. The most common bosons that we encounter in semiconductor physics are photons and phonons. Their number is not conserved: they may be created or destroyed. Argue why for them, $\mu = 0$. Therefore, for photon and phonons of energy $\hbar\omega$, the Bose–Einstein distribution simplifies to $f_{BE}(\mathcal{E}) = \frac{1}{e^{\frac{\hbar\omega}{k_b T}} - 1}$.

(4.5) The Boltzmann limit of switching

Electrons follow the Fermi–Dirac distribution in allowed energy bands of semiconductor based diodes and transistors that are used for digital logic circuits. It is experimentally observed that the current carried by these electrons under certain conditions depends on a voltage control V and temperature T as $I \propto e^{\frac{qV}{k_b T}}$. This is because the current is proportional to the number of electrons at energies much larger than the Fermi level E_F.

(a) Make a connection to the number of electrons available to carry current with the Fermi–Dirac integrals shown in Fig. 4.25.

(b) Show that irrespective of the orders of the Fermi–Dirac integrals, the number of electrons available to carry current changes as $e^{\frac{E_F}{k_bT}}$ when the Fermi energy is much lower than the minimum allowed electron energy.

(c) Assuming that the Fermi energy E_F can be directly changed with energy qV supplied from a battery of potential V, explain the experimental observations $I \propto e^{\frac{qV}{k_bT}}$.

(d) Show that to increase the current from I_1 to I_2, we must apply a voltage $V_2 - V_1 = \frac{k_bT}{q}\ln(\frac{I_2}{I_1})$.

(e) In particular, to increase the current by a factor of 10 at room temperature $T = 300$ K, show that we must apply a voltage of 60 mV. This is known as the **Boltzmann limit** for switching. In conventional transistors and diodes, this relation limits how abruptly one can switch diodes and transistors, and sets a lower limit of energy required to do useful operations with such devices.

(4.6) **The Joyce-Dixon approximation**
In this chapter we have seen how to obtain the carrier concentration n when the Fermi level E_F is known. For example, in 3D, the relation is $n = N_c F_{\frac{1}{2}}(\frac{E_F - E_c}{k_bT})$, where N_C is an effective band-edge density of states. Sometimes, we also need to know the inverse relation, meaning if we know the carrier density at a certain temperature, we wish to know where is the Fermi level.

(a) Prove the Joyce and Dixon result for 3D degenerate semiconductors, when $\eta = \frac{E_F - E_c}{k_bT} >> +1$ and E_c is the conduction band edge energy,

$$\frac{E_F - E_c}{k_bT} = \ln(\frac{n}{N_c}) + \frac{1}{\sqrt{8}}\frac{n}{N_c} - (\frac{3}{16} - \frac{\sqrt{3}}{9})(\frac{n}{N_c})^2 + ... \tag{4.34}$$

(b) Find the analogs of the 3D Joyce-Dixon result for 2D and 1D cases.

(4.7) **The Dulong–Petit law**
The specific heat capacity of a solid at high temperatures was known to follow the empirical Dulong–Petit law, for which classical thermodynamics provided an explanation. The Dulong–Petit law states that the specific heat capacity of a solid is $c_v = 3nk_b$, where n is the atomic density, and k_b the Boltzmann constant.

(a) Show that a mass M with a spring constant K when pulled a distance x from its equilibrium position has a total energy that is exactly twice the average kinetic and average potential energies.

(b) Assume the solid is composed of N atoms in volume V tied to each other through springs. Because of the mass-spring system, each atom is allowed to vibrate in three directions around their mean positions, and they also have potential energy due to the springs. Argue why the total energy absorbed by the solid at a temperature is $\mathcal{U} = 2 \cdot N \cdot \frac{3}{2}k_bT = 3Nk_bT$.

(c) Now show that the specific heat capacity $c_v = \frac{1}{V}\frac{\partial \mathcal{U}}{\partial T} = 3nk_b$, i.e., it is independent of temperature. Calculate the specific heat capacity for $n \sim 10^{23}/\text{cm}^3$ which is a typical density of atoms in a solid, and compare with experimentally measured values at high temperatures.

(d) Investigate when does the Dulong Petit law applies to solids, and under what conditions it breaks down. Find the typical specific heat capacities of semiconductors such as silicon and germanium, and compare with the prediction of the Dulong–Petit law. The deviations from this relation are due to the **quantization of lattice vibrations**, and is the topic of the next two problems.

(4.8) **Einstein's theory of heat capacity**
The experimentally measured specific hear capacity of solids in the late 19th century was found to be much smaller than the values the classical Dulong–Petit result predicted. And it was not universal either, it varied from solid to solid. In 1907, Einstein solved this mystery by applying the idea of quantization for the first time to vibrations in solids. This work is considered to be the first quantum theory of solids. In this problem, we re-do Einstein's derivation and in the process also gain an intimate acquaintance with heat and phonons in solids.

(a) Instead of Dulong–Petit's assumption of the classical equipartition of energy for N atomic oscillators, Einstein considered each atom is allowed a unique mechanical oscillation frequency ω_0. Drawing from Planck's prescription for the blackbody radiation (see Exercise 3.2), he postulated that the energies allowed for the mechanical vibration of the atom are not continuous, but *quantized* according to $E_n = n\hbar\omega_0$, where $n = 0, 1, 2, ...$ Show that in this case, assuming the Boltzmann statistics for occupation probability of energy $e^{-\frac{E_n}{k_b T}}$, the average energy for each atom is

$$\langle E \rangle = \frac{\hbar\omega_0}{e^{\frac{\hbar\omega_0}{k_b T}} - 1}. \tag{4.35}$$

(b) Show that because each atom can move in 3 directions, and there are N atoms in volume V where $n = N/V$, the total heat energy per unit volume stored in the atomic vibrations of the solid is

$$u_v = \frac{3N}{V}\langle E \rangle = 3nk_b \cdot \frac{(\hbar\omega_0)/k_b}{e^{\frac{\hbar\omega_0}{k_b T}} - 1}. \tag{4.36}$$

(c) The specific heat capacity is $c_v = \frac{\partial u_v}{\partial T}$. Make a plot of this quantity vs T and compare with Fig. 4.18. In particular, show why when $T >> \hbar\omega_0/k_b$, $c_v \to 3nk_b$, and the Dulong–Petit law is recovered. Also show why as $T \to 0$ K, $c_v \to 0$ as well. This is how Einstein solved the heat-capacity mystery.

(d) There is a direct analogy between Einstein's resolution of the specific heat anomaly and Planck's resolution of the ultraviolet catastrophe. Planck's resolution involved quanta of light (photons), and Einstein's quanta of mechanical vibrations in solids are called phonons. Draw the analogies between the two theories.

Walther Nernst, the person behind the third law of thermodynamics was so impressed by Einstein's explanation of the specific heat capacity that he travelled from Berlin to Zurich just to meet him in person. Nernst's fame and scientific clout is believed to have been instrumental in giving Einstein the first scientific break and high credibility.

(4.9) **Debye's theory of heat capacity**
In spite of the success of the Einstein theory, the quantitative behavior of the specific heat capacity of solids followed a T^3 dependence as $T \to 0$ K, and his theory did not explain this behavior. It also underestimated the value of c_v. This experimental fact was explained by a modification to Einstein's theory in 1912 by Peter Debye. Instead of assuming all atoms vibrate at the same frequency ω_0, Debye argued that atoms should be allowed to vibrate over a range of frequencies ω. The frequencies ω are related to the wavelength of sound waves carried by the solid, $\omega = v_s k$ where v_s is the speed of sound, and $k = 2\pi/\lambda$ where λ is the wavelength.

(a) By fitting of sound waves in a 3D solid of volume V and N atoms of atomic density $n = N/V$ and counting the total number of modes, show that the maximum energy of vibration is $\hbar\omega_D = \hbar v_s(6\pi^2 n)^{\frac{1}{3}}$. The frequency $\omega_D = 2\pi\nu_D = v_s(6\pi^2 n)^{\frac{1}{3}}$ is called the **Debye frequency** of vibrations of a solid. Estimate ν_D in a few solids and show it is in the range of 10^{12} Hz.

(b) By summing over the energies of the vibrational modes and using the density of states to go from a sum to an integral, show that the energy per unit volume is

$$u_v = p \cdot (3n) \cdot (k_b\theta_D) \cdot (\frac{T}{\theta_D})^4 \cdot \int_0^{\frac{\theta_D}{T}} dx \frac{x^3}{e^x - 1}, \tag{4.37}$$

where p is the number of modes of the sound wave. For a 3D solid, there is one longitudinal and two transverse polarizations for each mode, so $p = 3$.

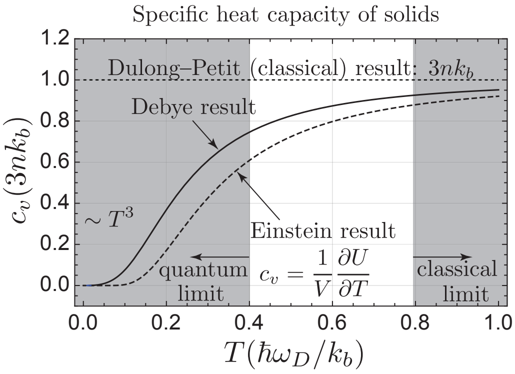


Fig. 4.18 The specific heat capacity c_v of solids as a function temperature. The classical Dulong–Petit result is reached only at high temperatures. Einstein first explained why $c_v \to 0$ as $T \to 0$ K by invoking quantization of lattice vibrations. Debye refined it to explain the T^3 variation of c_v as $T \to 0$ which was experimentally observed, but not explained in the Einstein theory. Both the quantum theories approach the classical Dulong–Petit law $c_v = 3nk_b$ at high temperatures.

(c) The derivative of Equation 4.37 with temperature gives the specific heat capacity $c_v = \frac{\partial u_v}{\partial T}$. Make a plot of the specific heat capacity as calculated by the Debye formula against the Einstein formula, and compare with Fig. 4.18.

(d) Show that as $T \to 0$ K, the Debye formula gives $c_v \to T^3$, which is consistent with experimental data.

Debye's vibrational modes with $\omega = v_s k$ are now identified and measured for solids as the **acoustic phonon** modes, and Einstein's $\omega = \omega_0$ constant modes are close to the **optical phonon** modes. They had no experimental means to identify and distinguish between these modes at that time. The Einstein and Debye formulations of the specific heat capacity of solids were the first applications of quantum mechanics to solids and matter. Note that Schrödinger and Heisenberg's formulation of quantum mechanics, the Bose–Einstein and Fermi–Dirac statistics, and harmonic oscillator problem of quantum mechanics only appeared more than a decade *after* the quantum theories of specific heat capacity, so none of those concepts are critical to truly explain the behavior of specific heat of solids!

(4.10) **Rotational and vibrational degrees of freedom**

When the heat capacity of a gas of N hydrogen *molecules* of density $n = N/V$ in a 3D box of volume V is measured as a function of temperature, the heat capacity per unit volume is found to go through steps as shown in Fig. 4.19: at the lowest temperatures it is $c_v = \frac{3}{2} n k_b$, which increases to $c_v = \frac{5}{2} n k_b$ above $T = \theta_{rot}$ and to $c_v = \frac{7}{2} n k_b$ above $T = \theta_{vib}$. This mysterious behavior was explained by quantum mechanics, and the equipartition of energy.

(a) Imagine each hydrogen molecule crudely as a point mass. Explain why because of translational degrees of freedom of this point mass, in a dilute concentration limit when the molecules are far apart from each other, the specific heat should be $c_v = \frac{\partial u_v}{\partial T} = \frac{3}{2} n k_b$.

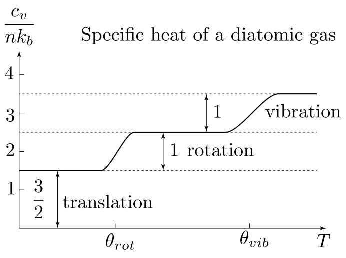


Fig. 4.19 Specific heat variation of a diatomic gas as a function of temperature.

(b) Now imagine each hydrogen molecule as a dumbbell. Then, argue why in addition to the three translational degrees of freedom, two degrees of freedom appear due to rotation. The molecule can rotate around the line joining the atoms, or around the perpendicular bisector of this line. If the distance between the atoms does not change, this is called a **rigid rotor**. From quantum mechanics, it is found that the rotational energy is quantized, taking values $\hbar\omega_{rot} l(l+1) = k_b \theta_{rot} l(l+1)$, where l is an integer, and θ_{rot} a characteristic temperature for rotation. Explain why the specific heat increases by $\frac{2}{2} n k_b$, and why this happens above $T = \theta_{rot}$

(c) Finally, the atoms in the molecular dumbbell can also vibrate with respect to each other. A 1-dimensional harmonic oscillator has 2 degrees of freedom as its state may be defined by the position and velocity. Argue then why the specific heat will further increase by an additional $\frac{2}{2} n k_b$ for $T > \theta_{vib}$.

(d) For hydrogen, experiments show that $\theta_{rot} \sim 85$ K, and $\theta_{vib} \approx 6332$ K. The gas can sometimes dissociate before such high temperatures are reached. There are further modes of energy such as electronic energies that can get excited at high temperatures. Do some research on this very high temperature behavior and compare with other gases.

(4.11) **Interacting fermions and bosons**

The Fermi–Dirac and Bose–Einstein distributions were derived under the assumption of non-interacting fermions and bosons. In the presence of interactions, the distributions functions are modified. Write a short paragraph on what are the modifications, and what are the experimental signatures of such modifications.

(4.12) **Useful integrals for magnetotransport**

Here we encounter two integrals related to the Fermi–Dirac function:

(a) Show that for $\beta = 1/(k_b T)$ and $f_0(E) = 1/(\exp[\beta(E-\mu)]+1)$ the Fermi–Dirac function,

$$\int_{-\infty}^{+\infty} dE \cdot \left(-\frac{\partial f_0}{\partial E}\right) \cdot e^{ikE} = \frac{\frac{\pi k}{\beta}}{\sinh\left(\frac{\pi k}{\beta}\right)}. \tag{4.38}$$

(b) Show that

$$\int_{-\infty}^{+\infty} dE \cdot \left(-\frac{\partial f_0}{\partial E}\right) \cdot \cos\left(2\pi s \frac{E}{\hbar\omega_c} - \frac{\pi}{4}\right) = \frac{2\pi s^2 \frac{k_b T}{\hbar\omega_c}}{\sinh\left(2\pi s^2 \frac{k_b T}{\hbar\omega_c}\right)} \cdot \cos\left(2\pi s \frac{\mu}{k_b T} - \frac{\pi}{4}\right). \tag{4.39}$$

These integrals will play an important role in quantum magnetotransport properties discussed in Chapter 25.

(4.13) **Stirling approximation**

A large part of statistical mechanics hinges on the Stirling approximation $\ln(N!) \approx N\ln(N) - N = N\ln(N/e)$ for large N. Here is one of many ways to derive the result.

(a) Plot $\ln(N!)$ and $N\ln(N) - N$ for $N = 1 - 10^5$ and see how good is the approximation.

(b) Consider the integral $\int_0^\infty dx e^{-\alpha x} = \frac{1}{\alpha}$. Differentiate both sides $(N+1)$ times with respect to α to prove that $\int_0^\infty dx x^N e^{-\alpha x} = \frac{N!}{\alpha^{N+1}}$. Putting $\alpha = 1$ gives an integral representation of $N!$.

(c) Show that the integrand $x^N e^{-x}$ with $\alpha = 1$ peaks at $x_o = N$. Prove that as N becomes large, the peak is very strong.

(d) Prove that if a function $f(x)$ is sharply peaked at $x = x_0$, then the integral $I = \int dx e^{f(x)}$ may be approximated by $I \approx e^{f(x_0)}$, if $x = x_n = n$ is the integer n, and $\Delta x = x_{n+1} - x_n = 1$.

(e) Prove the Stirling formula using the above results, by writing $x^N e^{-x} = e^{N\ln(x) - x}$.

(4.14) **Entropy and limits of solar cells**

The concepts of entropy and the Fermi–Dirac and Bose–Einstein distributions determine the maximum efficiency with which semiconductor solar cells can convert energy from light to electricity. These concepts are discussed in detail in Chapter 28. In this exercise, we appreciate these limits qualitatively.

(a) Make a plot of the Planck distribution of photon number versus the photon energy at the temperature $T = 5800$ K of the sun's surface. This is roughly the spectrum that reaches the outer atmosphere.

(b) Suppose a semiconductor of bandgap $E_g = 1.0$ eV is used for making a solar cell. It does not absorb photons below its bandgap. Then calculate the fraction of the photons that can be absorbed by the solar cell.

(c) Once the energy is converted from light to the energy of electrons and holes in the semiconductor, the energy distribution changes to a Fermi–Dirac function, which rapidly thermalizes with the temperature of the semiconductor atoms. A thermodynamic Carnot-type efficiency limit is then also defined based on the temperature of the sun and the semiconductor, as discussed in Chapter 28. Try to argue on heuristic grounds what this limit should be, and check the actual quantitative answer in Chapter 28.

Electrons in the Quantum World

5

5.1 In Schrödinger equation we trust 84
5.2 The free electron 85
5.3 Not so free: Particle on a ring 87
5.4 The electron steps into a higher dimension: 2D 98
5.5 Electrons in a 3D box 105
5.6 The particle in a box 111
5.7 The Dirac delta potential 113
5.8 The harmonic oscillator 114
5.9 The hydrogen atom 114
5.10 Chapter summary section 115
Further reading 116
Exercises 116

In this chapter we subject the electron to the laws of quantum mechanics embodied by the Schrödinger equation, and quantum statistics embodied by the Pauli exclusion principle as indicated in Fig. 5.1. We discover that by upgrading classical mechanics and thermodynamics to their quantum versions, we can explain a vast array of experimental facts for which the classical Drude model of the electron failed.

The goals of this chapter are to explore:

- How does application of quantum mechanics and statistics to free electrons help reveal their wavefunctions ψ, momentum, energy, density of states, energy density, and quantum mechanical currents?
- How do the above physical properties of electrons depend on the number of dimensions the electron is allowed to move in? Specifically, what are the similarities and differences of the properties for electrons moving in 1D, 2D, and 3D?
- How did quantum mechanics and quantum statistics of free electrons resolve the discrepancies of the classical Drude model of the behavior of electrons in solids, and what phenomena could still not be explained?
- A few exactly solved situations when electrons are not free, but form **bound** states due to potentials that trap them.

In the process, we encounter a few *exactly* solved problems of quantum mechanics. The set of exactly solved problems is precious, because they form a rigorous underpinning on which rests the edifice of condensed matter physics and the physics of semiconductor nanostructures. We introduce the techniques to find the physical dynamics of *single* electrons – such as its momentum, energy, and the current it carries – all now in the significantly updated quantum version where the wave-nature is imposed on the electron from the beginning. Then we find that because of the Pauli exclusion principle, many-electron systems have an ”internal” energy that is both very large and very bewildering, because it has simply no counterpart in classical mechanics. We introduce the concept of the *density of states*, and our bewilderment turns to joy as we realize that we can not only explain the failures of the

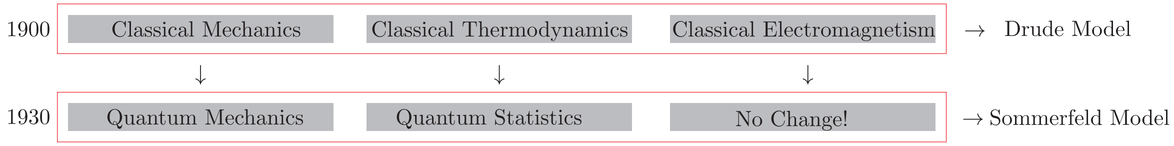


Fig. 5.1 The major changes in the understanding of the electronic properties of solids involved quantum-facelifts of mechanics and statistics, but no changes in electromagnetism were necessary. The new quantum mechanical rules that govern the behavior of the electron are the point of discussion of this chapter.

Drude model, but have discovered a powerful new bag of tricks that explain and *predict* a far richer range of physical behavior of electrons in bulk materials *and* in nanostructures for which classical mechanics and thermodynamics are insufficient.

5.1 In Schrödinger equation we trust

As we discussed in Chapter 3, **all physically measurable information** about the quantum states of the electron are buried in the state vector $|\psi\rangle$. By projecting the state vector to the real space we get the wavefunction $\psi(x) = \langle x|\psi\rangle$. In this chapter, we learn how to extract useful information from $\psi(x)$ by applying the corresponding operators of physical observables on them. To do that, we have to first solve the time-independent Schrödinger equation for an electron in various potentials $V(x)$:

$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x) + V(x)\psi(x) = E\psi(x). \tag{5.1}$$

The set of solutions $\langle x|n\rangle = \psi_n(x)$ will then be the eigenfunctions corresponding to states of *definite energy* with corresponding eigenvalues E_n. As we learnt in Chapter 3, the **states of definite energy are also stationary states**. They form the most convenient basis for describing the situations when the potential deviates from the ideal, i.e., if $V(x) \to V(x) + W(x,t)$. Thus, the states of definite energy form the basis to uncover what happens when we perturb the quantum system.

In the next sections we will examine the behavior of electrons in a few potentials $V(x)$ that are central to understand the electronic properties of solids in general, and semiconductor physics in particular. For these potentials, we repeat the following procedure to extract physical information:

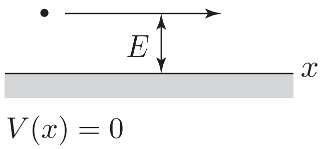
- solve the Schrödinger equation *exactly* to obtain the wavefunction $\psi(x)$ [1],
- the allowed momentum p_x,
- the allowed energy eigenvalues E,
- the density of states $g(E)$,

---

[1] The Schrödinger equation can be solved exactly in an analytical form only for a very few potentials. We will cover most of them in this chapter. For general potentials, numerical solutions may be obtained but the computational cost may be high.

- the *total* energy $\mathcal{U}$, *average* energy u, and energy density u_v of many electrons, and
- the quantum mechanical current J.

We begin with the simplest of potentials: when $V(x) = 0$.



5.2 The free electron

For $V(x) = 0$, the Schrödinger equation reads

$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x) = E\psi(x). \tag{5.2}$$

The equation has the most general solution of the form

$$\boxed{\psi(x) = Ae^{ikx} + Be^{-ikx},} \tag{5.3}$$

where

$$\boxed{k = \sqrt{\frac{2m_eE}{\hbar^2}} = \frac{2\pi}{\lambda}.} \tag{5.4}$$

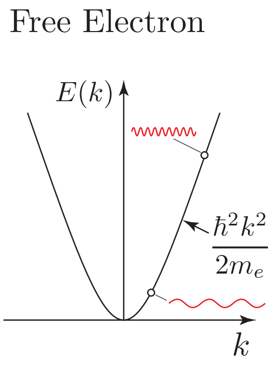


Fig. 5.2 Free electron in 1D. The energy dispersion is parabolic, and *all* k and all $E > 0$ are allowed.

We emphasize that the allowed wavelengths λ can take *any* value. Thus, the allowed k values are *continuous* for $V(x) = 0$. The allowed energy eigenvalues, also continuous, are

$$\boxed{E(k) = \frac{\hbar^2k^2}{2m_e}.} \tag{5.5}$$

Fig. 5.2 shows this parabolic energy dispersion of the free electron. It would not be a stretch to say that this simplest energy dispersion is also one of the most important in all of condensed matter physics. The **curvature of the parabola is inversely proportional to the mass** of the quantum particle. If the particle is heavy, the energies are lower, you can imagine the parabola being pulled down by the heavy mass. Later we will see that the allowed energy eigenvalues of the electron in a semiconductor crystal will develop bands and gaps because of the formation of standing electron waves. Even so, within each electron band, the energy dispersion will again assume a parabolic form at the band edges, but with different *effective* masses than the free electron mass because of the presence of a periodic crystal potential.

We note that the general solution in Eq. 5.3 represents a superposition of two waves: one going to the right ($\psi_{\rightarrow}(x) = Ae^{ikx}$) and the other to the left ($\psi_{\leftarrow}(x) = Be^{-ikx}$). Since it is a "mixed" state, clearly it is *not* a state of a definite momentum. We verify this by operating upon the wavefunction by the momentum operator:

$$\begin{aligned}\hat{p}_x\psi(x) &= -i\hbar\frac{d}{dx}\psi(x) = -i\hbar(ikAe^{ikx} - ikBe^{-ikx}) \\ &= \hbar k(Ae^{ikx} - Be^{-ikx}) \neq p\psi(x)\end{aligned} \tag{5.6}$$

but for just the right-going state we get

$$\hat{p}_x\psi_{\rightarrow}(x) = -i\hbar\frac{d}{dx}\psi_{\rightarrow}(x) = -i\hbar(ikAe^{ikx}) = \hbar k\psi_{\rightarrow}(x) = p\psi_{\rightarrow}(x) \tag{5.7}$$

which *is* a state of definite momentum $\hbar k$. For a right-going momentum eigenstate $|+k\rangle$, whose wavefunction is $\psi_{\rightarrow}(x) = Ae^{ikx}$, we find that the quantum charge current density upon using Equation 3.28 is

$$J(+k) = \frac{q}{2m_e}(\psi^{\star}_{\rightarrow}\hat{p}_x\psi_{\rightarrow} - \psi_{\rightarrow}\hat{p}_x\psi^{\star}_{\rightarrow}) \implies \boxed{J(+k) = q|A|^2\frac{\hbar k}{m_e}.} \tag{5.8}$$

Note that the units are in amps, because $|A|^2$ has units of 1/length. Similarly, for a left-going state $|-k\rangle$ with wavefunction $\psi_{\leftarrow}(x) = Be^{-ikx}$, the charge current density is

$$J(-k) = \frac{q}{2m_e}(\psi^{\star}_{\leftarrow}\hat{p}_x\psi_{\leftarrow} - \psi_{\leftarrow}\hat{p}_x\psi^{\star}_{\leftarrow}) = -q|B|^2\frac{\hbar k}{m_e}. \tag{5.9}$$

From an analogy to the "classical" charge current density $J = qnv$, where $n \sim |A|^2$ or $n \sim |B|^2$ is the particle density[2], we identify that the state $|+k\rangle$ has a velocity $\frac{\hbar k}{m_e}$, and the mirror-reflected state $|-k\rangle$ has a velocity $-\frac{\hbar k}{m_e}$. The net current due to the right- and left-going states then is given by

$$J_{net} = J(+k) + J(-k) = q\cdot(|A|^2 - |B|^2)\cdot\frac{\hbar k}{m_e}, \tag{5.10}$$

which states that the net current going to the right is the difference of the two currents. If we now had a composite state $\psi(x) = C_1e^{ik_1x} + C_2e^{ik_2x} + ... = \sum_n C_ne^{ik_nx}$, the net current is easily found as $J_{net} = q\sum_n |C_n|^2\frac{\hbar k_n}{m_e}$.

In classical mechanics, a particle of mass m has a kinetic energy $E = p^2/(2m)$, and a velocity $v = p/m = dE/dp$. In quantum mechanics, the particle has a wave-like nature. By analogy we cautiously[3] define the **group velocity** of a quantum particle as proportional to the *slope* of the energy dispersion curve $E(k)$

$$\boxed{\mathbf{v}_g(\mathbf{k}) = \nabla_{\mathbf{p}}E(\mathbf{p}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k}).} \tag{5.11}$$

This definition will assume an increased importance when the $E(k)$ dependence is different from that of a free electron; in Chapter 9 we will see that for electrons in a periodic crystal, this result remains exact, with the meaning of $\mathbf{k}$ being modified to that of a crystal momentum. Using the group velocity, we can write the charge current as $J(k) = qv_g(k)f(k)$, where $f(k)$ is the occupation probability of state $|k\rangle$. Using this procedure, we can find the quantum charge current carried by any superposition state $|\psi\rangle = \sum_k A_k|k\rangle$ we can cook up.

The free electron wavefunction *cannot* be normalized, because it extends over all space from $-\infty \leq x \leq +\infty$. Physical quantities that have to do with density, such as the "carrier density", "density of states" or the "energy density" are ill defined for the completely free electron because of the infinite volume it lives in. To normalize it, we wrap the infinitely long line and join the infinities to form a circle. So we first put the electron in a circular ring to calculate these quantities.

---

[2] If you are uncomfortable with this statement, I am with you. $n \sim |A|^2$ is true *only if* the particle is confined, as we will see in the next section. The completely free electron wavefunction is not normalizable!

[3] A more modern formulation of the velocity of a quantum state is $\mathbf{v}_g(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k}) + \frac{q}{\hbar}\mathbf{E}\times\mathcal{B}(\mathbf{k})$, where the second term appears due to an effective magnetic field in the momentum (or $\mathbf{k}$-space). Here $\mathcal{B}(k)$ is a Berry curvature of bands and $\mathbf{E}$ is the electric field. The Berry curvature of free electrons, and *most* semiconductor bands is zero. The second term causes currents perpendicular to the direction of the electric field under certain special cases.

5.3 Not so free: Particle on a ring

Wavefunctions and Momenta: Fig. 5.3 (a) shows an electron restricted to move on a circular ring of circumference L, with $V(x) = 0$. Though it is not exactly a 1D problem, we assign one linear coordinate x to the particle's location. We demand all solutions to the Schrödinger equation to be *single-valued* functions of x. Because the loop closes on itself[4], the electron wavefunction must satisfy

$$\psi(x+L) = \psi(x) \to e^{ik(x+L)} = e^{ikx} \to e^{ikL} = 1 \to kL = 2n\pi \tag{5.12}$$

to be single-valued. This is only possible if

$$\boxed{k_n = \frac{2\pi}{L}n,} \text{ where } n = 0, \pm 1, \pm 2, ... \tag{5.13}$$

where $\psi_n(x) = Ae^{ik_n x}$. We at once see that for the particle on a ring, the set of allowed k_n are *discrete* as indicated in Fig. 5.2, and thus the allowed values of the momentum are discrete:

$$\boxed{p_n = \hbar k_n = \frac{h}{2\pi}\frac{2\pi}{L}n = n\frac{h}{L},} \tag{5.14}$$

that is, the allowed values of the electron momentum are *quantized*. The smallest spacing of the allowed wavevectors is precisely

$$\boxed{\Delta k = k_{n+1} - k_n = \frac{2\pi}{L}.} \tag{5.15}$$

Because the angular momentum is $\mathbf{L} = \mathbf{r} \times \mathbf{p}$, we find that

$$\mathbf{L}_n = \mathbf{r} \times \mathbf{p} = \frac{L}{2\pi} \times \hbar k_n \hat{\mathbf{z}} = \frac{L}{2\pi} \times \frac{2\pi\hbar}{L} n\hat{\mathbf{z}} \implies \boxed{L_n = n\hbar,} \tag{5.16}$$

i.e., like the linear momentum, the *angular* momentum of the electron on a ring is also quantized, and can only take values equal to integer multiples of Planck's constant $..., -2\hbar, -\hbar, 0, +\hbar, +2\hbar, ...$ We gain a physical intuition of what Planck's constant $\hbar$ actually means – it is a measure of the angular momentum of a quantum particle. For example, if I tie a 1 kg mass to a 1 m string and spin it at 1 m/s radial velocity, the angular momentum is $L_{cl} = 1$ J·s. So for this classical situation, I will be providing the mass $n = L_{cl}/\hbar \sim 10^{34}$ quanta of angular momenta – and I may feel like Superman in the quantum world.

---

[4] This periodic boundary condition is referred to as the Born von-Karman boundary condition. It is mathematically distinct from the "hard-wall" boundary condition that we impose for the particle in a box problem in Section 5.6, but the physics of the interior will not be affected by this choice. Exercise 5.4 makes this point clear. Furthermore, the circular 'ring' geometry is an artificial construct for 1D confined electrons in a straight quantum wire. It serves to capture the physics with least mathematical complications. It is also no more artificial if a 1D wire is indeed wound into a ring.

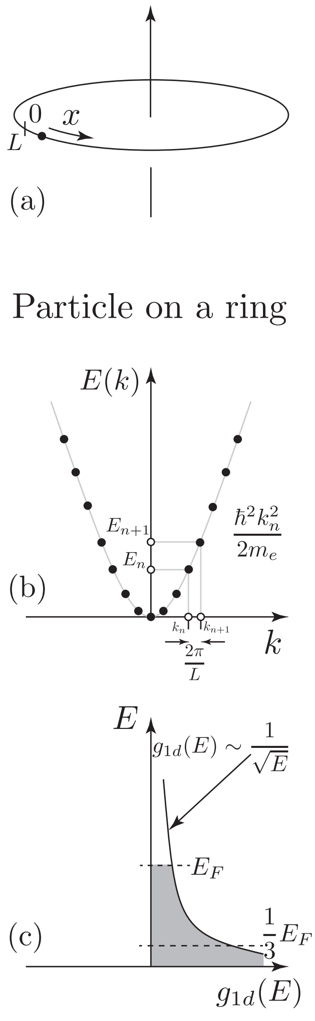


Fig. 5.3 Putting the electron on a ring *quantizes* the allowed wavevectors k_n, and as a result the momentum, the angular momentum, and the energy of the particle are quantized. The density of states for the electron on a ring with parabolic energy dispersion goes as $1/\sqrt{E}$, counting how the allowed eigenvalues distribute in energy if we were to put many electrons on the ring.

But what this example really tells us is precisely *how small a quantum of angular momentum actually is*!

As promised, unlike the free electron case, the eigenfunctions of the particle on a ring can be normalized:

$$\int_0^L dx|\psi_n(x)|^2 = 1 \to |A|^2 \times L = 1 \to A = \frac{1}{\sqrt{L}} \to \boxed{\psi_n(x) = \frac{1}{\sqrt{L}}e^{ik_n x}.} \qquad (5.17)$$

Note that $n = 0$ is *allowed* as a result of the periodic boundary condition, and the probability density of this state $|\psi_0(x)|^2 = 1/L$ is a constant in x, as is the probability of every state of definite momentum. We observe that the set of functions $[..., \psi_{n-1}(x), \psi_n(x), \psi_{n+1}(x), ...]$ are mutually orthogonal because

$$\begin{aligned}\langle m|n\rangle = \int_0^L dx \langle m|x\rangle\langle x|n\rangle &= \int_0^L dx \psi_m^\star(x)\psi_n(x) \\ &= \int_0^L dx \frac{e^{i\frac{2\pi}{L}(n-m)x}}{L} = \delta_{n,m}.\end{aligned} \qquad (5.18)$$

Note this integration carefully, because it is a very important one, *and a simple one*, yet it remains an endless source of confusion to beginners. If $n \neq m$, the integral is $\int_0^L dx \cdot e^{i2\pi N(\frac{x}{L})}$ where $N = n - m$ is an integer. The value of the integral is zero because we are integrating $\cos(2\pi N \frac{x}{L}) + i\sin(2\pi N\frac{x}{L})$ over an integer number of periods as $x : 0 \to L$, with equal positive and negative parts. But when $n = m$, $e^{i\frac{2\pi}{L}(n-m)x} = e^0 = 1$, and the integral is simply $\int_0^L \frac{dx}{L} = 1$. Because the integral is 1 when $n = m$ and it is $= 0$ when $n \neq m$, we write it compactly using the Kronecker delta symbol $\delta_{n,m}$. Note that the Kronecker delta function[5] takes two discrete arguments n, m, and can only assume two values: 0 and 1. Couldn't be any simpler – fear not the Kronecker delta $\delta_{n,m}$ – it is your friend!

We also make a note that the set of linearly independent functions $\psi_n(x)$ is *complete*, meaning if you give me *any* smooth function $f(x)$ in 1D, I can write the function as $f(x) = \sum_n c_n\psi_n(x)$ with suitable coefficients c_n. This is what we mean by linear superposition: the amazing thing about quantum mechanics is the electron on the ring is allowed to be in *any* state that is a linear superposition of the eigenstates.

Energy and Density of States: The allowed energy eigenvalues are

$$\boxed{E(k_n) = \frac{\hbar^2 k_n^2}{2m_e} = n^2\frac{(2\pi\hbar)^2}{2m_e L^2} = n^2\frac{h^2}{2m_e L^2} = E_n.} \qquad (5.19)$$

The energy eigenvalues are also *quantized*, and grow as n^2 as seen in Fig. 5.3 (b). Because the electron is allowed to be in the $n = 0$ state, the minimum energy allowed is $E = 0$. This will *not* be the case if we put the particle in a box in Section 5.6. Two important physical intuitions we should take away from this example are:

---

[5] The Kronecker delta function is different from the Dirac delta function $\delta(x)$, which takes a continuous set of arguments x, and can assume any value between 0 and ∞.

- The smaller the circle, the larger the allowed energies ($L \downarrow \implies E_n \uparrow$), and
- The smaller the mass, the larger the allowed energies ($m \downarrow \implies E_n \uparrow$).

The first statement is one of *quantum confinement*: the smaller the space we fit a quantum particle into, the larger will be its energy. This is because the wavelength must become very small to fit in the space, which means high $k = 2\pi/\lambda$, and high energy $E = (\hbar k)^2/(2m_e)$.

A glance at the density of energy eigenvalues along the energy (y)-axis in Fig. 5.3 shows that they are more densely spaced at low energies, and become sparse at higher energies[6]. We can guess that the dependence with energy must go as $1/E^{\gamma}$, where $\gamma > 0$. To find the 1D density of states quantitatively, we note that between $k \to k + dk$, there are $\frac{dk}{\frac{2\pi}{L}}$ allowed states. For completeness, for each allowed k-state we introduce a *spin degeneracy* g_s, which is $g_s = 2$ for free electrons of up and down spins, and a *valley degeneracy* g_v which is the number of copies of such parabolic dispersions that may be present in the k-space[7]. The total state density $G_{1d}(E)$ in energy is then

$$g_s g_v \frac{2dk}{\frac{2\pi}{L}} = G_{1d}(E)dE \implies g_{1d}(E) = \frac{G_{1d}(E)}{L} = \frac{2g_s g_v}{2\pi\frac{dE}{dk}}$$

$$\implies \boxed{g_{1d}(E) = \frac{g_s g_v}{2\pi}\left(\frac{2m_e}{\hbar^2}\right)^{\frac{1}{2}}\frac{1}{\sqrt{E}}.} \qquad (5.20)$$

Because in the 1D ring, the electron could be moving clockwise or counterclockwise, we use $2dk$ to account for the two dk's for $+k$ and $-k$ wavevectors. We note that the 1D DOS decreases as $1/\sqrt{E}$, and has a singularity as $E \to 0$.

Now instead of a single electron, if we fill the ring with N electrons, what would be the total energy of the ensemble? If there are N electrons in the ring, their 1D density is $n_{1d} = N/L$. We will first completely neglect[8] the Coulomb interaction between the electrons and assume they are non-interacting. Though this sounds like heresy, let me assure you that it is actually OK[9]. Because electrons are fermions, the moment we go from one electron to two, the Pauli exclusion principle kicks in. Let us first address a subtle point about the physics of many fermions head-on.

An essential point of clarification now is that if there are N electrons on a ring, we have a quantum *many particle* state, not a *single particle* state. A *single particle* quantum state that is in a superposition of many eigenstates is represented by a *sum* wavefunction that is a linear superposition of the form $\psi(x) = \sum_k c_k \psi_k(x)$, where $\sum_k |c_k|^2 = 1$ ensures there indeed is a single, or 1 particle. But the wavefunction of a fermionic *many particle* quantum state is an anstisymmetrized *product* of single particle wavefunctions of the form $\psi(x_1, x_2, ..., x_N) = \sum_p (-1)^p \psi_{k_1}(x_1)\psi_{k_2}(x_2)...\psi_{k_N}(x_N)/\sqrt{N}$, where p is a permutation of x_n's and k_m's, compactly written as **Slater determinants**. We encountered this fundamental property of fermions in Chapter 3, Equation 3.34. The specific case of the 2-electron state has a wavefunction $\psi(x_1, x_2) = (\psi_{k_1}(x_1)\psi_{k_2}(x_2) - \psi_{k_1}(x_2)\psi_{k_2}(x_1))/\sqrt{2}$. The Pauli exclusion principle is baked into this antisymmetric wavefunction.

---

[6] The allowed energy eigenvalues are *equally* spaced along the x- or k-axis since $\Delta k = \frac{2\pi}{L}$.

[7] For free electrons $g_v = 1$. In crystals there may be several equivalent energy valleys due to crystal symmetries, when $g_v \neq 1$. We will encounter them later, but here we retain g_v for generality. Also for free electrons $g_s = 2$, but this degeneracy is broken in a magnetic field, or in magnetic materials as discussed in later chapters.

[8] In crystalline solids, electrons can experience bizarre situations when they not only not repel each other, but attract! This counterintuitive mechanism is responsible for superconductivity.

[9] The justification came from Lev Landau, and goes under the umbrella of what is called the Fermi-liquid theory. We will encounter it at a later point. Besides, the discussion here is equally applicable to neutrons – which are fermions that actually do not have charge.

Though the many particle electron wavefunction looks like a mess, the fact that they are non-interacting saves us from mathematical complications. **The total momentum, energy, and current of the many particle distribution then becomes the sum of the momenta, energies, or currents of the single particle states $\psi_k(x)$ it is made of.** Chapter 4 also gave us the occupation function of the single particle state k at thermal equilibrium of the non-interacting many particle state after enforcing the Pauli exclusion principle as the simple result:

The equilibrium electron occupation function of a distribution of a large number of non-interacting electrons follows the Fermi–Dirac distribution $f(E_k)$.

Armed with these powerful simplifying results for many electrons, let us look at the $T = 0$ K situation, when $f(E) = 1$ for $0 \leq E \leq E_F$ and $f(E) = 0$ for $E > E_F$. The electrons then must fill up to a Fermi wavevector k_F and Fermi level E_F as shown in Fig. 5.3 such that[10]

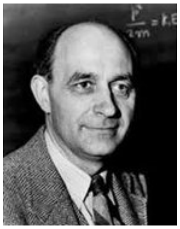


Fig. 5.4 Enrico Fermi, a towering figure whose contributions extended into all areas of physics, both as a theorist and an experimentalist. Nobel prize recipient in 1938. Fermi and Dirac formulated the quantum statistics of particles that follow the exclusion principle - such particles are named fermions after him. He played a key role in nuclear physics, and led the team that produced an artificial chain reaction for the first time.

$$g_s g_v \times \frac{2k_F}{\frac{2\pi}{L}} = N \implies \boxed{k_F = \frac{\pi}{2g_s g_v} n_{1d}} \implies \boxed{E_F^{1d}(0\text{ K}) = \frac{\hbar^2 \pi^2 n_{1d}^2}{2g_s^2 g_v^2 m_e}}. \tag{5.21}$$

This is a remarkable result: the *ground state* of the electron ensemble at $T = 0$ K already has a large amount of energy. For example, if $n_{1d} \sim 10^8$/cm typical for a metal, then for $g_s = 2, g_v = 1$ the electron states with the highest energy have $\lambda_F = 2\pi/k_F \sim 0.4$ nm and $E_F \sim 10$ eV. This is the energy picked up by an electron in a 10 Volt potential [11], or if we were to provide this energy in the form of heat, $k_b T = E_F$ would lead to $T \sim 10^5$ K! **Where did all this energy come from?**

The root of this energy reserve is the fermionic nature of the electron, and the Pauli exclusion principle. There is simply no classical explanation of this energy, it is of a pure quantum origin. We will shortly see that the very high conductivity of metals even at the lowest temperatures is a direct result of these internal quantum energy reserve. And the electrons at the highest energy are *fast*: the Fermi velocity is $v_F = \frac{\hbar k_F}{m_e} = \frac{h n_{1d}}{4m_e} \sim 5 \times 10^7$ cm/s. This is a typical Fermi velocity in metals and degenerately doped semiconductors: a Fermi velocity of $v_F \sim 10^8$ cm/s is roughly $v_F \approx c/300$, that is, the highest energy electrons can be considered to be whizzing around at 1/300 times the speed of light even at $T = 0$ K due to the Pauli exclusion principle!

The **Fermi energy** $E_F^{1d}(0\text{ K}) = \hbar^2\pi^2 n_{1d}^2/8m_e$ in Equation 5.21 is a characteristic value at $T = 0$ K, also called the **chemical potential** (see Exercise 4.4) with a symbol $\mu_{1d} = E_F^{1d}(T = 0\text{ K})$. The carrier density at

[10] Note that the number (or density) of electrons completely defines the Fermi wavevector k_F and the Fermi energy E_F at $T = 0$ K. This holds for all dimensions, and for all $E(k)$ relations.

[11] If we pack even more fermions in small volumes, the Fermi energy is much larger. For example, because of the tight packing of fermions – protons and neutrons – in the nucleus of an atom, the Fermi energy reaches ~ 40 MeV – yes, *Mega* eV!

a temperature T depends on the **Fermi level** $E_F^{1d}(T)$:

$$n_{1d}(T) = \frac{g_s g_v}{L} \sum_k f(k) = \frac{g_s g_v}{L} \int_{-\infty}^{+\infty} \frac{dk}{\frac{2\pi}{L}} \frac{1}{1 + e^{\frac{\frac{\hbar^2 k^2}{2m_e} - E_F^{1d}(T)}{k_b T}}}. \qquad (5.22)$$

The dimensionless variables $u = (\frac{\hbar^2 k^2}{2m_e})/k_b T$ and $\eta = E_F^{1d}(T)/k_b T$ convert the carrier density at any temperature to

$$n_{1d}(T) = \underbrace{g_s g_v (\frac{2\pi m_e k_b T}{h^2})^{\frac{1}{2}}}_{N_c^{1d}} F_{-\frac{1}{2}}(\frac{E_F}{k_b T}) = N_c^{1d} F_{-\frac{1}{2}}(\eta), \qquad (5.23)$$

where we have made use of the definition of Fermi–Dirac integrals $F_j(\eta)$ from Chapter 4, Equation 4.25. We have collected the coefficient into N_c^{1d} in a form that is called the **band-edge density of states**, a name that can (and will!) be justified only later for semiconductors. We will see shortly that the relation between the free electron carrier density and temperature in the most general form for d dimensions is

$$\boxed{n_d = N_c^d F_{\frac{d-2}{2}}(\eta),} \text{ where } \boxed{N_c^d = g_s g_v (\frac{2\pi m_e k_b T}{h^2})^{\frac{d}{2}}} \qquad (5.24)$$

is the band-edge DOS for d dimensions. The order of the Fermi–Dirac integral is $(d-2)/2$, which is $-1/2$ for $d = 1$ or 1D, 0 for 2D, and $1/2$ for 3D. Thus Equation 5.23 is the special case for 1D of Equation 5.24. Fig. 5.5 shows the value of the band edge density of states for free electrons for $d = 1, 2, 3$ for temperatures ranging from $1 - 1000$ K. Though we write N_c^d to avoid clutter, the temperature dependence should not be forgotten. At room temperature, the free electron band-edge DOS are: for 1D, $N_c^{1d}(300\text{K}) = 4.6 \times 10^6/\text{cm}$, for 2D, $N_c^{2d}(300\text{K}) = 10^{13}/\text{cm}^2$, and for 3D, $N_c^{3d}(300\text{K}) = 2.5 \times 10^{19}/\text{cm}^3$. These densities typically mark the crossover from classical behavior for $n_d << N_c^d$ to quantum behavior for $n_{1d} >> N_c^d$.

Now if the total density of electrons n_{1d} does not change[12] with temperature, then the Fermi level $E_F^{1d}(T)$ must adjust to ensure that

$$n_{1d}(T = 0\text{ K}) = n_{1d}(T) \implies \boxed{n_{1d} = N_c^{1d} F_{-\frac{1}{2}}(\eta).} \qquad (5.25)$$

Solving the boxed equation gives the temperature dependence of the Fermi level $E_F^{1d}(T) = \eta \cdot k_b T$. No analytical expression is possible in 1D for the most general temperature. But if the temperature is such that $k_b T << \mu_{1d}$, then using $\eta = E_F^{1d}(T)/k_b T >> +1$, or the degenerate limit of the Fermi–Dirac integral given in Table 4.2, we obtain after some algebra, the relation

$$E_F^{1d}(T) \approx \mu_{1d}[1 + \frac{\pi^2}{12}(\frac{k_b T}{\mu_{1d}})^{\frac{5}{2}}] \text{ for } k_b T << \mu_{1d}. \qquad (5.26)$$

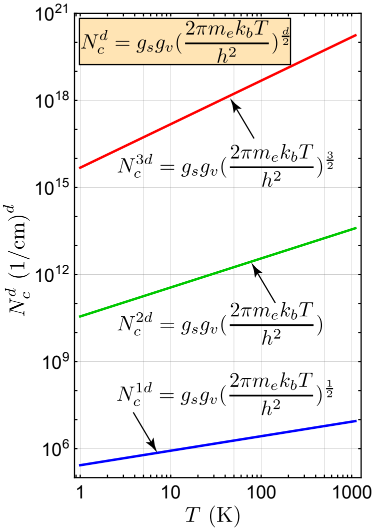


Fig. 5.5 The free electron band edge DOS N_c^d for $d = 1, 2, 3$ in units of 1/cm for 1D, 1/cm^2 for 2D, and 1/cm^3 for 3D for $g_s = 2, g_v = 1$. The physical significance of this characteristic electron density N_c^d is the following: if the electron concentration $n_d = N_c^d F_{\frac{d-2}{2}}(\eta)$ with $\eta = E_F/k_b T$ is much smaller than N_c^d, $F_{\frac{d-2}{2}}(\eta) << 1$, which implies $F_{\frac{d-2}{2}}(\eta) \approx e^\eta << 1$, which is possible if $E_F << 0$. This is a non-degenerate electron concentration, and most of its properties are close to classical. But if the electron concentration is much larger than N_c^d, $F_{\frac{d-2}{2}}(\eta) >> +1$, and $F_{\frac{d-2}{2}}(\eta) \approx \eta^{\frac{d}{2}}/\Gamma(\frac{d}{2}+1)$, and the Fermi level $E_F >> 0$. Most properties of this electron distribution are quantum mechanical and strongly affected by the exclusion principle.

[12] The carrier density in a band is typically fixed for metals as a function of temperature, which is the subject of interest in this chapter. For semiconductors, due to the proximity to an energy gap, the density can change with temperature. This we will encounter in later chapters.

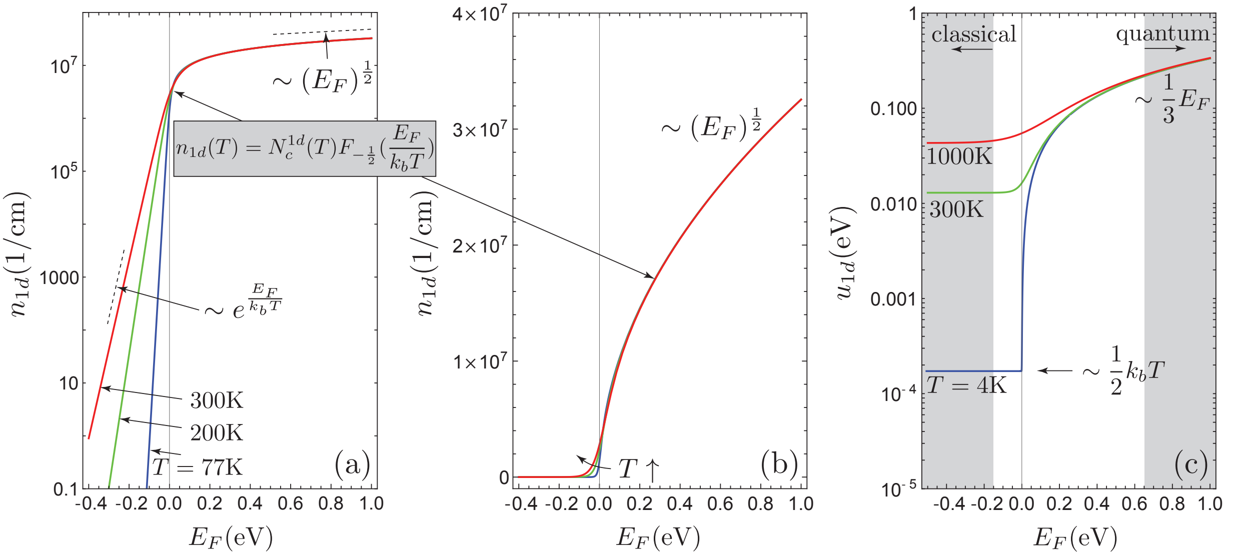


Fig. 5.6 Carrier statistics and energies for the 1D free electron gas. (a) The 1D electron density as a function of the Fermi level E_F governed by $n_{1d} = N_c^{1D} F_{-1/2}(E_F/k_b T)$ shown for various temperatures. For high carrier densities (the degenerate limit), the carrier density does not depend on T and depends on E_F as a polynomial power. For low densities (the non-degenerate limit), the dependence on E_F and T is $e^{E_F/k_b T}$, i.e., exponential and very sensitive both to the Fermi level and the temperature. (b) Shows the dependence in the linear scale. It is important to note that though it appears so, $n_{1d} \neq 0$ for $E_F << 0$, but is very small as can be seen in (a). (c) The average energy u_{1d} per electron of the 1D electron gas, which in the non-degenerate limit is the same as the classical equipartition of energy $\frac{1}{2}k_b T$, and in the high density quantum limit is $u_{1d} \sim \frac{1}{3}E_F$ at all temperatures.

In the other limit when $\eta = E_F^{1d}(T)/k_b T << -1$, also called the non-degenerate or the high temperature limit, $F_j(\eta) \approx e^{\eta}$, and we obtain

$$n_{1d} \approx N_c^{1d} e^{\frac{E_F^{1d}(T)}{k_b T}} \implies E_F^{1d}(T) \approx k_b T \ln(\frac{n_{1d}}{N_c^{1d}}). \tag{5.27}$$

The dependence of the 1D electron gas density on the Fermi level and the temperature are shown in Figs. 5.6 (a) and (b). Typical 1D carrier densities in metals are in the $n_{1d} \sim 10^7 - 10^8/\text{cm}$ range. In 1D semiconductors n_{1d} can reach $10^6/\text{cm}$, and can be reduced by several orders of magnitude by field-effect from a gate electrode.

Energy and Specific Heat: The total energy $\mathcal{U}$ of the ensemble of electrons of density n_{1d} at $T = 0$ K, and the average energy $u_{1d} = \mathcal{U}/N$ are then given by

$$\mathcal{U} = \int_0^{\mu_{1d}} dE \cdot E \cdot G_{1d}(E) \implies \boxed{u_{1d} = \frac{\mathcal{U}}{N} = \frac{\int_0^{\mu_{1d}} dE \cdot E \cdot G_{1d}(E)}{\int_0^{\mu_{1d}} dE \cdot G_{1d}(E)} = \frac{1}{3}\mu_{1d},} \tag{5.28}$$

and the energy density per length $u_v = \frac{\mathcal{U}}{L}$ is then given by

$$\boxed{u_v(1d) = \frac{1}{3} n_{1d}\mu_{1d}.} \tag{5.29}$$

That the average energy of the 1D electron distribution $\frac{1}{3}\mu_{1d}$ is less than $\frac{1}{2}\mu_{1d}$ is easily understood from the shape of the DOS $g_{1d}(E)$ in Fig. 5.3 (c), which is weighted heavier for lower energies.

When the 1D electron system is heated to a temperature T, the net energy stored in the electronic system increases from its $T = 0$ K value[13]. To find the increase, we can use the result from Chapter 4 Equation 4.33 on the quantum mechanical equipartition of energy for d dimensions, $u_d(T) = \frac{d}{2}k_bT\frac{F_{\frac{d}{2}}(\eta)}{F_{\frac{d-2}{2}}(\eta)}$. For $d = 1$ the average energy of the 1D electron gas is

$$u_{1d}(T) = \frac{1}{2}k_bT\frac{F_{\frac{1}{2}}(\eta)}{F_{-\frac{1}{2}}(\eta)} \approx \frac{1}{3}\mu_{1d} + \frac{2\pi^2}{9}\frac{(k_bT)^2}{\mu_{1d}}, \tag{5.30}$$

where the approximation is valid for $\eta >> +1$, or for temperatures such that $k_bT << \mu_{1d}$. The approximation requires some algebra and the use of Equations 5.26 with the approximations of the Fermi Dirac integrals given in Table 4.2. The average energy for 1D electrons is shown in Fig. 5.6 (c). When the carrier density is low, the average energy is the classical equipartition of energy result, with $k_bT/2$ energy per electron, and is linearly dependent on temperature. At high densities, the average energy becomes *independent* of temperature, and the quantum mechanical result takes over, with the average energy of $E_F/3 \approx \mu_{1d}/3$ dependent on the internal quantum energy reserve due to the exclusion principle. This very weak dependence of the energy per electron on temperature translates to a very small electronic specific heat compared to the classical prediction, as described next.

The energy density $u_v(1d)$ per unit length is then $u_v(1d) = n_{1d}u_{1d}(T)$. The specific heat of the electron system is defined as the energy intake per degree rise in the temperature, or

$$c_v^{1d} = \frac{\partial u_v(1d)}{\partial T} = \frac{4\pi^2}{9}(\frac{k_bT}{\mu_{1d}}) \cdot n_{1d}k_b. \tag{5.31}$$

In the *classical* equipartition of energy, each particle in 1D can absorb $\frac{1}{2}k_bT$ of energy, and the classical electronic specific heat in 1D is

$$c_v^{1d}(\text{classical}) = \frac{\partial u_v(1d)(\text{classical})}{\partial T} = \frac{\partial}{\partial T}(\frac{1}{2}n_{1d}k_bT) = \frac{1}{2}n_{1d}k_b. \tag{5.32}$$

So the quantum mechanical electronic specific heat is much smaller than the classical:

$$\frac{c_v^{1d}(\text{quantum})}{c_v^{1d}(\text{classical})} \approx \frac{\frac{4\pi^2}{9}(\frac{k_bT}{\mu_{1d}}) \cdot n_{1d}k_b}{\frac{1}{2}n_{1d}k_b} \approx \frac{8\pi^2}{9}(\frac{k_bT}{\mu_{1d}}) << 1. \tag{5.33}$$

The electronic specific heat predicted by classical thermodynamics and classical mechanics $\frac{1}{2}n_{1d}k_b$ is inconsistent with experiments. What is experimentally measured is about $1/100$ lower, depending on the temperature. Equation 5.33 for the first time explained the mystery of the very low electronic specific heats compared to the Drude model. We finally have a quantitative resolution of what was alluded to in Chapter 2, Fig. 2.9. When electrons in 1D are subjected to quantum mechanics (instead of classical mechanics), and to quantum statistics (instead of classical statistics, see Fig. 5.1), the much lower electronic specific heat is easily understood. Electrons far below the Fermi energy μ_{1d} cannot absorb any energy as all available energy states $E(k)$ up to μ_{1d} are occupied, and the Pauli exclusion principle forbids a k-state from occupation more than $g_s = 2$. On the other hand, electron k-states in a small slice of energies $k_b T$ around μ_{1d} can absorb energy and go to nearby empty states. Thus, only a fraction $\approx \frac{k_b T}{\mu_{1d}}$ of the total number of electrons are capable of absorbing heat. Since typical metals have $\mu_{1d} \approx 5$ eV, and $k_b T \approx 26$ meV at room temperature, $\frac{k_b T}{\mu_{1d}} \approx \frac{1}{200}$ explains why the experimental electronic specific heat is so small. This explanation was first provided by Arnold Sommerfeld – not for 1D, but for the 3D problem that we will see shortly.

---

[13] The physical meaning of "heated to a temperature T" is precise. It means that the electrons in the 1D system are in equilibrium with a heat bath of temperature T, and the Fermi–Dirac distribution has the temperature T in it.

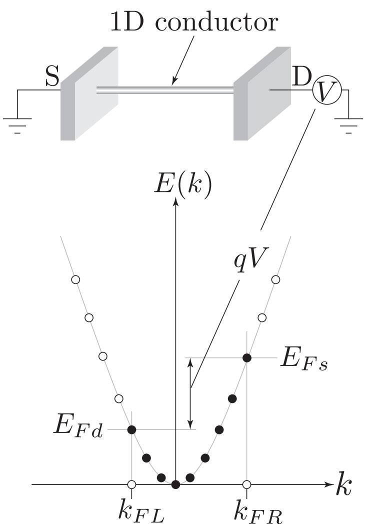


Fig. 5.7 Quantum picture of current flow in the 1D k-space due to a difference in the occupation functions of right-going and left-going eigenstates. The left contact is called the source, and the right contact the drain.

Quantum Current: In Fig. 5.7, a 1D conductor is connected to a source contact on the left, and a drain contact on the right which is at a voltage V. As discussed in Chapter 4, Section 4.8, if these contacts are ohmic, then the right-going electrons are in equilibrium with the left contact – the source – and are filled up to a quasi Fermi level E_{Fs}. The left-going electrons are in equilibrium with the drain contact, and are filled to a Fermi level E_{Fd}. The battery splits the source and the drain Fermi levels by $E_{Fs} - E_{Fd} = qV$. Let us define the dimensionless parameters $\eta_s = \frac{E_{Fs}}{k_b T}$, $\eta_d = \frac{E_{Fd}}{k_b T}$, and $v_d = \frac{qV}{k_b T}$ to keep the equations compact. We see that $\eta_s - \eta_d = v_d$. An ohmic contact does not change the total electron density in the 1D conductor, which implies that the total 1D electron density is the sum of the left and right going states, i.e., $n_{1d} = \frac{1}{2} N_c^{1d} [F_{-\frac{1}{2}}(\eta_s) + F_{-\frac{1}{2}}(\eta_d)]$. Thus, by solving the two equations

$$\boxed{\eta_s - \eta_d = v_d, \text{ and } n_{1d} = \frac{1}{2} N_c^{1d} [F_{-\frac{1}{2}}(\eta_s) + F_{-\frac{1}{2}}(\eta_d)]} \tag{5.34}$$

simultaneously, we can determine the unknowns η_s and η_d for any drain voltage V. As a sanity check, when $V = 0$, we recover $\eta_s = \eta_d = \eta$ and $n_{1d} = N_c^{1d}(T) F_{-\frac{1}{2}}(\eta)$.

We have used a particle on a ring for the analysis till this point, but the 1D conductor in Fig. 5.7 is not circular. Electrons on a truly closed circular ring can be ”contacted” by a time-varying magnetic flux by Faraday’s law $V = -\frac{d\Phi_m}{dt}$, but we do not consider such a case here and stick to a DC voltage V across a linear conductor. In exercise 5.4, it is shown that the periodic boundary condition and a hard-wall boundary condition lead to identical results for bulk states.

We now evaluate the quantum current flowing in the 1D wire. The net quantum mechanical current J_{1d}^R flowing to the right in the $+k$ direction when the Fermi level is at E_{Fs} is obtained by summing the current from all states with +ve k:

$$J^R_{1d} = q\frac{g_s g_v}{L}\sum_{k\geq 0} v_g(k)f(k) = q\frac{g_s g_v}{L}\int_0^\infty \frac{dk}{\frac{2\pi}{L}}\frac{1}{\hbar}\frac{dE(k)}{dk}f(k)$$

$$\implies J^R_{1d} = \frac{q g_s g_v}{2\pi\hbar}(k_b T)F_0(\eta_s). \tag{5.35}$$

Similarly, because the left-going electron states are in equilibrium with the drain electrode, the net quantum current flowing to the left is

$$J^L_{1d} = \frac{q g_s g_v}{2\pi\hbar}(k_b T)F_0(\eta_d). \tag{5.36}$$

Since $F_0(\eta) = \ln(1+e^{\eta})$, the net quantum current at any temperature for a 1D conductor is

$$\boxed{J_{1d} = J^R_{1d} - J^L_{1d} = \frac{q g_s g_v}{2\pi\hbar}(k_b T)[F_0(\eta_s) - F_0(\eta_d)] = \frac{q g_s g_v}{2\pi\hbar}(k_b T)\ln(\frac{1+e^{\eta_s}}{1+e^{\eta_d}}).} \tag{5.37}$$

Because for any voltage V we know η_s and η_d from Equation 5.34, the quantum mechanical current in a 1D conductor is completely solved for all voltages at all temperatures by Equation 5.37. But we must realize that this is a ballistic result, meaning we have assumed that there is no scattering in the 1D conductor. Scattering will lead to a reduction of the net current, and is the subject of later chapters in the book. Before we discuss the characteristics of the 1D quantum current, we point out that just like the general expression for the band-edge DOS N_c^d for d-spatial dimensions, an expression J^d for ballistic current exists. To see this, we write the 1D ballistic current as[14]

$$J_{1d} = \frac{q^2}{h}\cdot N_c^{0d}\cdot\frac{k_b T}{q}\cdot[F_0(\frac{E_{Fs}-E_c}{k_b T}) - F_0(\frac{E_{Fs}-E_c-qV}{k_b T})], \tag{5.38}$$

where $N_c^{0d} = g_s g_v = g_s g_v(\frac{2\pi m_e k_b T}{h^2})^0$ is the band-edge DOS of one dimension lower, of a 0D quantum dot which does not have bands, but discrete energy states. The general expression for the quantum current for free electrons moving in d-dimensions in space is

$$\boxed{J_d = \frac{q^2}{h}\cdot N_c^{d-1}\cdot\frac{k_b T}{q}\cdot[F_{\frac{d-1}{2}}(\frac{E_{Fs}-E_c}{k_b T}) - F_{\frac{d-1}{2}}(\frac{E_{Fs}-E_c-qV}{k_b T})].} \tag{5.39}$$

Let us make some physical sense of the results for 1D, and see some very non-classical features of the quantum mechanical current. For temperatures low enough such that $\eta_s >> +1$ and $\eta_d >> +1$, $\ln(1+e^{\eta}) \approx \eta$, and the quantum mechanical 1D current is

$$J_{1d} \approx \frac{q g_s g_v}{2\pi\hbar}(k_b T)(\eta_s - \eta_d) = g_s g_v\frac{q^2}{h}V \implies \boxed{G_{1d} = \frac{J_{1d}}{V} = g_s g_v\frac{q^2}{h},} \tag{5.40}$$

where we find that the quantum conductance of a 1D wire depends only on the fundamental constants q, h, and **does not depend on the**

[14] In this expression, we introduce the energy level E_c as the minimum allowed electron energy of the problem at the source injection point. For the free electron, $E_c = 0$, whereas in a metal or a semiconductor, it will become the bottom of the conduction band energy.

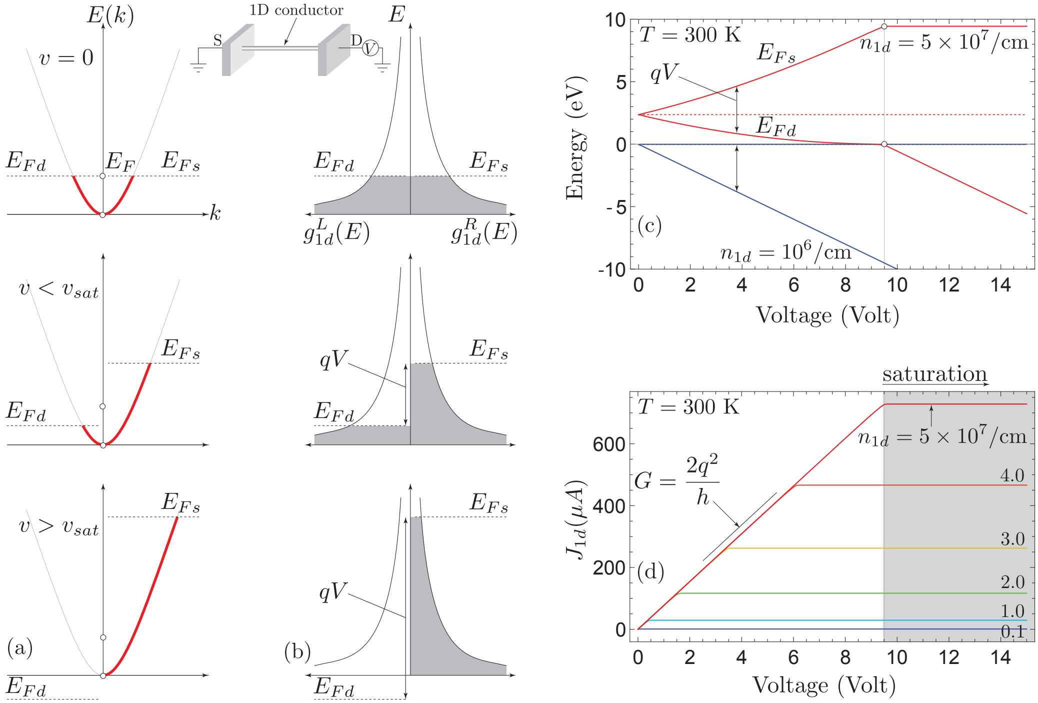


Fig. 5.8 Quantum current in a 1D conductor. (a) $E(k)$ figures show the changes in the corresponding occupied electron states for $n_{1d} = 5 \times 10^7$/cm for three different voltages. (b) The right-going and left-going density of states, and how their occupation is changed with the applied voltage. (c) The calculated Fermi level $E_F = \eta_F \cdot k_b T$ at $V = 0$ shown as a dashed line, and the split normalized Fermi levels $E_F = \eta_s \cdot k_b T$ and $E_{Fd} = \eta_d \cdot k_b T$ for nonzero normalized voltages $qV = v_d \cdot k_b T$ for two values of 1D electron density at 300 K. The red curves are for $n_{1d} = 5 \times 10^7$/cm, and the blue for $n_{1d} = 10^6$/cm. (d) The resulting quantum mechanical current flowing in response to the voltage for six values of 1D electron densities ranging from $0.1 - 5.0 \times 10^7$/cm. For example, at a 1D electron density of $n_{1d} = 10^7$/cm, the maximum (or saturation) current is ∼ 70 μA.

length of the wire! For $g_s = 2$ and $g_v = 1$, the quantum of conductance G_{1d} corresponds to a resistance $1/G_{1d} \approx \frac{h}{2q^2} \approx 12,950\ \Omega$. This value of conductance is experimentally achieved at high carrier densities in the ballistic limit when there is no scattering of the electron states. This quantum value of conductance has been experimentally observed in nanowires, carbon nanotubes, and in "break junctions". The first surprise then is that the quantized conductance does not decrease with increasing length, which is the classical Ohm's law, explained by the Drude model in Chapter 2. The classical resistance increases as the length, but the quantum current is ballistic: an electron wave incident from the source contact makes it all the way to the drain if there is no scattering. That scattering is necessary to explain the classical result of Ohm's law will be discussed in later chapters.

Evaluating the quantum current further for a range of 1D electron densities and voltages reveals further new surprises that are inconsistent with classical Drude-model type current flow, yet are consistent with what is experimentally measured for nanostructures. To calculate the total current, we use Equation 5.37, with Equation 5.34. In Fig. 5.8, panels (a) and (b) schematically show the occupation of the free electron energies in the k-space and the density of states picture. The right-going carriers are in equilibrium with the source and the left-going carriers in equilibrium with the drain.

Panel (c) of Fig. 5.8 shows the calculated values of $E_{Fs} = \eta_s \cdot k_b T$ and $E_{Fd} = \eta_d \cdot k_b T$ for two starting 1D electron densities at $T = 300$ K. The split is equal to the applied voltage, as evaluated from Equation 5.34. At $T = 300$ K, and for a "low" electron density $n_{1d} = 10^6/\text{cm}$ shown in blue, η_F is slightly negative. Upon application of the drain voltage, the two Fermi levels split, but η_s stays close to η_F, whereas η_d is lowered by $v_d = qV/k_b T$. On the other hand, for a high 1D electron density of $n_{1d} = 5 \times 10^7/\text{cm}$ shown in red, $\eta_F \approx +90$ and splits initially nearly symmetrically into η_s and η_d around η_F for low voltages. As the voltage is increased, the split becomes increasingly asymmetric, till η_d hits zero at $v_d = v_{sat}$. For larger voltages, η_s flattens, and η_d nosedives into the negative, picking up any increase in the voltage, all the while maintaining $\eta_s - \eta_d = v_d$. Panel (d) of Fig. 5.8 shows the quantum or ballistic current calculated from Equation 5.37 for several electron densities.

For each 1D electron density, there is a characteristic drain voltage V_{sat} beyond which the η_s becomes flat, and the η_d dives into the negative. The net current saturates[15] at this voltage and higher, since almost all available electrons are moving to the right, as indicated in the $E(k)$ diagram. When this happens, the 1D conductor is not able to provide any more current; any further increase in voltage goes into decreasing the left-going carriers, but there are almost none of them, and their contribution to the current is negligible. For voltages less than V_{sat}, the 1D ballistic conductance is the quantum of conductance $2q^2/h$ for all electron densities. An expression for V_{sat} may be obtained from a simple geometric analysis of Fig. 5.8. The onset of current saturation occurs when $E_{Fd} = 0$. Because $E_{Fs} - E_{Fd} = qV$, $V_{sat} = E_{Fs}^{sat}/q$ at the voltage when $E_{Fd} = 0$. Writing the Fermi level at $V = 0$ as E_F, since all electrons are moving to the right at saturation, and in d-dimensions $E_F \sim k_F^2 \sim n_d^{2/d}$,

$$E_{Fd}^{sat} = E_F \cdot 2^{\frac{2}{d}} \implies \boxed{V_{sat} = \frac{E_F \cdot 2^{\frac{2}{d}}}{q}.} \tag{5.41}$$

For example, for 1D electrons of density $n_{1d} = 5 \times 10^7/\text{cm}$, $E_F \sim 2.4$ eV, and the saturation voltage is $V_{sat} = 2^2 \cdot E_F/q \sim 9.4$ Volt, consistent with the top curve in Fig. 5.8 (d). The saturation voltage decreases with the electron density monotonically[16].

What is the meaning of a negative quasi Fermi level? For free electrons, the lowest allowed electron energy is $E(k = 0) = 0$, and cannot be negative. However, the Fermi level characterizing the Fermi–Dirac occupation function can be negative. For example, E_{Fd} characterizing the drain Fermi level, which is equal to the Fermi level of the left-going

[15] The saturation of current in a conductor is an extremely important characteristic for several applications. Saturation makes the output current independent of the voltage drop across the conductor, as long as it is larger than V_{sat}. This means that one conductor can be used to drive several similar conductors, and the voltage drop of these driven conductors may fluctuate and not affect the current in the driver. This is called fan-out, and is of utmost importance in electronic circuits.

[16] The dependence of the saturation voltage on the electron densities in various dimensions go as $V_{sat} \propto n_{1d}^2, n_{2d}^1, n_{3d}^{\frac{2}{3}}$.

carriers can go below $E(k) = 0$. In the real-space points along the conductor where this condition is met, there are very few left-going states filled, all electrons are essentially moving to the right. In later chapters, when we discuss not the free electron case, but electrons in semiconductors, the lowest allowed electron energy in the conduction band edge will be E_c, and the Fermi level will be measured with respect to this energy: $\eta_d = (E_{Fd} - E_c - qV)/k_bT$.

This quantum $J_{1d} - V$ behavior of a 1D electron gas has three non-classical features, not explainable by the Drude model, and are not consistent with Ohm's law[17]. They are:

1) In the linear regime, the **1D conductance is $2q^2/h$, and does not depend on the length of the conductor**. This we have pointed out earlier.

2) In addition to not depending on the length of the conductor, the 1D quantum conductance is also **independent of the 1D electron density!** In the Drude model, a higher electron density leads to a higher conductance, but not necessarily so in the quantum limit.

3) The fact that the quantum current saturates is non-classical, and not seen in Ohm's law. The **saturation current depends on the 1D electron density**, even though the 1D conductance in the linear regime does not. So the quantum picture sets an upper limit on the current a 1D conductor can carry.

Looking closer at the J_{1d} vs v_d plot in Fig. 5.8, we recognize that the current-voltage behavior looks very much like the output characteristics of a transistor. We realize that if by some means we could change the 1D electron density with an extra, third electrode, without drawing any current into this electrode, we can get all the curves of this plot with the same conductor. This indeed is the **underlying working principle of a transistor**, about which we discuss in later chapters.

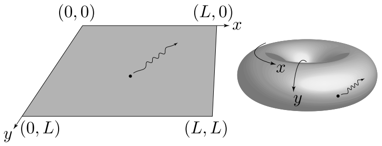


Fig. 5.9 Periodic boundary conditions in 2D leads to a torus.

5.4 The electron steps into a higher dimension: 2D

Wavefunctions and Momenta: Now let the electron move in two dimensions – say in a square box of side L and area $A = L^2$ shown in Fig. 5.9. The spatial coordinate is now the 2D vector $\mathbf{r} = (x, y)$, and the wavevector is $\mathbf{k} = (k_x, k_y)$. Similar to the 1D case, by subjecting the electron to the Schrödinger equation in two dimensions with $V(\mathbf{r}) = 0$, we find that the allowed wavefunctions are

$$\boxed{\psi(\mathbf{r}) = \frac{1}{\sqrt{L^2}} e^{i(k_x x + k_y y)} = \frac{1}{\sqrt{A}} e^{i\mathbf{k}\cdot\mathbf{r}}.} \tag{5.42}$$

If we paste the edges along $x = 0$ and $x = L$, and then the edges along $y = 0$ and $y = L$, the periodic boundary condition in real space and in $\mathbf{k}$-space leads to a torus (Fig. 5.9). This requires $e^{ik_xL} = 1$ and $e^{ik_yL} = 1$, which yields the allowed wavevectors and momenta

---

[17] We have discussed several aspects of the quantum 1D conductor in detail. Our hard work will now pay off handsomely as we step to higher dimensions, because the 2D and 3D cases are qualitatively similar, except for some quantitative details. What is remarkable is that almost all the discussions and expressions we have described here for the *free* electrons carry over with minor changes to metals and semiconductors when electrons experience the periodic potential of atoms. So spend enough time understanding the physics of the 1D free electron gas, and re-read this section if some parts are not clear, till you are comfortable with both the qualitative and quantitative aspects of the 1D electron gas.

$$\boxed{\mathbf{k} = (k_{n_x}, k_{n_y}) = \frac{2\pi}{L}(n_x, n_y)} \implies \boxed{\mathbf{p} = \hbar\mathbf{k}, |\mathbf{p}| = \frac{h}{L}\sqrt{n_x^2 + n_y^2},} \tag{5.43}$$

where n_x, n_y are *independent* integers ..., $-2, -1, 0, 1, 2, \ldots$. In the $\mathbf{k}$ space, the allowed set of points form a rectangular grid, each of area $(\frac{2\pi}{L})^2$. Each point in this grid defines an allowed state for electrons. This is shown in Fig. 5.10 (a).

Energy and Density of States: The allowed energy of each such allowed $\mathbf{k}$-point defines the paraboloid

$$\boxed{E(k_x, k_y) = \frac{\hbar^2}{2m_e}(k_{n_x}^2 + k_{n_y}^2) = E(n_x, n_y) = (n_x^2 + n_y^2)\frac{h^2}{2m_eL^2} = \frac{\hbar^2|\mathbf{k}|^2}{2m_e}.} \tag{5.44}$$

To find the 2D DOS $g_{2d}(E)$, we try our intuition like in the 1D case, but we must be careful. Indeed the energies bunch up as we approach $(k_x, k_y) = (0, 0)$, but we must not forget that unlike the 1D case where there were mere two points of equal energy, we have an entire *circle* of equal-energy states in 2D as shown in Fig. 5.10 (a) on the surface of the paraboloid. In the 2D case, the number of states increases as k^2 as the area of the $\mathbf{k}$-space circle, as does the energy. Since the DOS is the number of states per unit energy, the two effects cancel out. The resulting 2D DOS is *independent* of electron energy:

$$g_sg_v\frac{2\pi k dk}{(\frac{2\pi}{L})^2} = G_{2d}(E)dE \implies \frac{G_{2d}(E)}{L^2} = \boxed{g_{2d}(E) = \frac{g_sg_vm_e}{2\pi\hbar^2}\Theta(E).} \tag{5.45}$$

Here $2\pi k dk$ is the area of the thin ring of thickness dk around the circle of radius k as indicated in Fig. 5.10 (a). Because each state occupies an area $(\frac{2\pi}{L})^2$ in the $\mathbf{k}$-space, there are $\frac{2\pi k dk}{(\frac{2\pi}{L})^2}$ energy states in the ring, from which we get the DOS $g_{2d}(E)$. We also note that because for a free electron, $g_s = 2$ and $g_v = 1$, the 2D DOS is typically written as $g_{2d}(E) = \frac{m_e}{\pi\hbar^2}$ for $E > 0$. The DOS is shown in Fig. 5.10 (b).

The fact that the 2D DOS for a parabolic energy dispersion is a constant in energy plays a very important role in semiconductor field-effect transistors, where the conducting channel is a 2D electron gas. Moving to the many-electron picture in 2D, let us put N non interacting electrons in the area A so that their density per unit area is $n_{2d} = N/A$. At $T = 0$ K, we apply the Pauli exclusion principle again and find that we must fill the states from the center $\mathbf{k} = (0, 0)$ to a sharply defined **Fermi circle** of radius k_F given by

$$g_sg_v\frac{\pi k_F^2}{(\frac{2\pi}{L})^2} = N \implies \boxed{k_F = \sqrt{\frac{4\pi n_{2d}}{g_sg_v}}.} \tag{5.46}$$

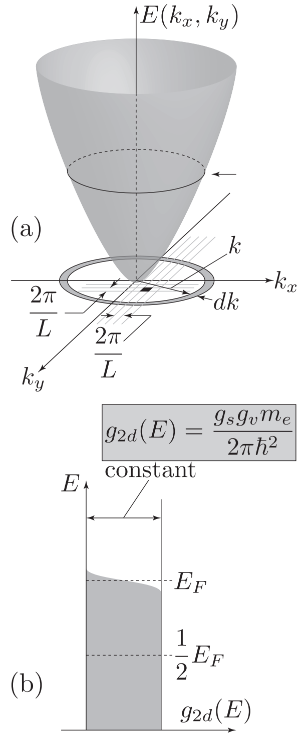


Fig. 5.10 (a) Energy eigenvalues on a paraboloid $E(k_x, k_y)$ indicating the 2D grid in the k-space, and (b) the density of states (DOS) for free electrons in two dimensions.

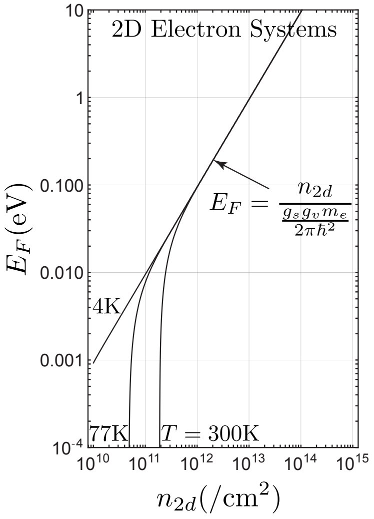


Fig. 5.11 Dependence of the Fermi level on the density of free 2D electrons and on temperature. At high densities and low temperatures, $E_F \approx n/(g_{2d})$.

For $g_s = 2$ and $g_v = 1$, the Fermi wavevector is $k_F = \sqrt{2\pi n_{2d}}$. In semiconductor field-effect transistors, typical sheet densities of 2D electron gases (2DEGs) are $n_{2d} \sim 10^{12}/\mathrm{cm}^2$. The Fermi wavevector is then $k_F \sim 2.5 \times 10^8/\mathrm{m}$, implying a Fermi wavelength of $\lambda_F = \frac{2\pi}{k_F} \sim 25$ nm. On the other hand, for a 2D metal with $n_{2d} \sim 10^{16}/\mathrm{cm}^2$, the Fermi wavelength is much shorter, $\lambda_F \sim 0.25$ nm, a typical interatomic distance in a metal or semiconductor crystal. At nonzero temperatures, the 2D electron density is

$$n_{2d} = \frac{g_s g_v}{L^2} \sum_k f(k) = \frac{g_s g_v}{L^2} \int_0^\infty \frac{2\pi k dk}{(\frac{2\pi}{L})^2} \frac{1}{1 + e^{\frac{\frac{\hbar^2 k^2}{2m_e} - E_F^{2d}(T)}{k_b T}}}. \tag{5.47}$$

By making the change of variables $u = \frac{\hbar^2 k^2}{2m_e}/k_b T$ and $\eta = E_F^{2D}(T)/k_b T$ and using the definition of the Fermi–Dirac integral $F_0(\eta) = \ln(1 + e^\eta)$, we rewrite the temperature-dependent 2D electron density as

$$n_{2d}(T) = N_c^{2d} \cdot F_0(\eta) = \underbrace{\frac{g_s g_v m_e}{2\pi\hbar^2} k_b T}_{N_c^{2d}} \cdot \ln(1 + e^{\frac{E_F^{2d}(T)}{k_b T}}), \tag{5.48}$$

where we have defined the 2D band-edge DOS $N_c^{2d} = \frac{g_s g_v m_e}{2\pi\hbar^2} k_b T$. Using $\hbar = h/2\pi$ and rewriting it as $N_c^{2d} = g_s g_v (\frac{2\pi m_e k_b T}{h^2})$, we see that the free electron band-edge DOS and the electron density follows the d-dimensional rules that $N_c^d = g_s g_v (\frac{2\pi m_e k_b T}{h^2})^{\frac{d}{2}}$ and $n_d = N_c^d F_{\frac{d-2}{2}}(\eta)$, with $d = 2$, as was stated in Equation 5.24.

For non-zero temperatures, the smearing of the Fermi–Dirac distribution near $E = E_F$ makes the Fermi circle diffuse. To get the electron density in a 2DEG at any temperature T, we use the 2D DOS and use the fact that the electron density does not change with temperature to get the Fermi level E_F thus:

$$n_{2d} = \int_0^\infty dE \cdot g_{2d}(E) \cdot f(E) = N_c^{2d} \ln(1 + e^{\frac{E_F}{k_b T}})$$

$$\implies \boxed{E_F = k_b T \ln(e^{\frac{n_{2d}}{N_c^{2d}}} - 1).} \tag{5.49}$$

If the temperature is low enough so that $n_{2d} >> N_c^{2d}$, the Fermi level is given simply by $E_F \sim k_b T \cdot \frac{n_{2d}}{N_c^{2d}} \sim \frac{n_{2d}}{g_{2d}}$, or $n_{2d} = g_{2d} E_F$, where $g_{2d} = g_s g_v \frac{m_e}{2\pi\hbar^2}$. The meaning of this relation: electron density = (DOS) × (Energy window). The dependence of the Fermi level on temperature and the 2D free electron density is shown in Fig. 5.11. As an example, typical 2DEGs in transistors have carrier concentrations $n_{2d} \sim 10^{12} - 10^{13}/\mathrm{cm}^2$. At low temperatures, and even at room temperature, these 2DEGs are well in the quantum regime and the Fermi level has a weak dependence on temperature, if at all. Fig. 5.12 (a, b) show the dependence of the 2D electron gas density on the Fermi

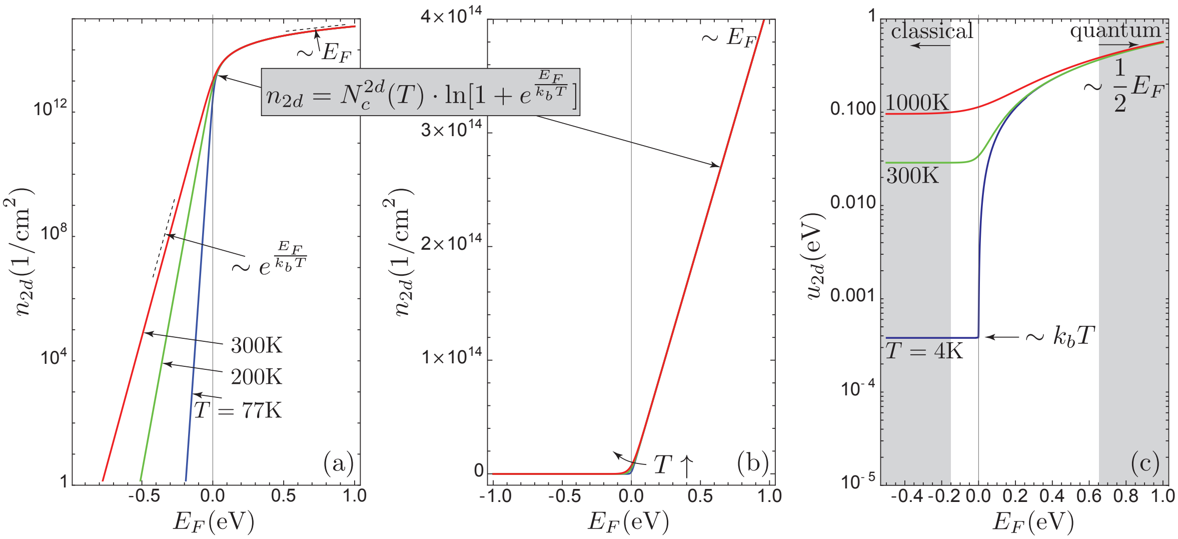


Fig. 5.12 Carrier statistics and energies for the 2D free electron gas. (a) The 2D electron density as a function of the Fermi level E_F governed by $n_{2d} = N_c^{2d} F_0(E_F/k_bT)$ shown for various temperatures. For high carrier densities (the degenerate limit), the carrier density does not depend on T and depends on E_F as a polynomial power. For low densities (the non-degenerate limit), the dependence on E_F and T is e^{E_F/k_bT}, i.e., exponential and very sensitive both to the Fermi level and the temperature. (b) Shows the dependence in the linear scale. It is important to note that though it appears so, $n_{2d} \neq 0$ for $E_F << 0$, but is very small as can be seen in (a). (c) The average energy u_{2d} per electron of the 2D electron gas, which in the non-degenerate limit is the same as the classical equipartition of energy $2 \times \frac{1}{2}k_bT = k_bT$, and in the high density quantum limit is $u_{2d} \sim \frac{1}{2}E_F$ at all temperatures.

level and temperature. The exponential dependence of n_{2d} on E_F when E_F is negative is not visible clearly in the linear plot in Fig. 5.12 (b), but unmistakable in the log plot in Fig. 5.12 (a). This exponential dependence of electron density on potential is important in the off-state characteristics of transistors and diodes.

Part (c) of Fig. 5.12 shows the average energy per electron at three temperatures, for different Fermi levels. At a high 2D electron gas density, we can obtain the total energy $\mathcal{U}$ and the average electron energy u_{2d} at $T = 0$ K as

$$\mathcal{U} = \int_0^\infty dE \cdot E \cdot G_{2d}(E) \cdot f(E)$$

$$\implies \boxed{u_{2d} = \frac{\mathcal{U}}{N} = \frac{\int_0^\infty dE \cdot E \cdot G_{2d}(E) \cdot f(E)}{\int_0^\infty dE \cdot G_{2d}(E) \cdot f(E)} = \frac{1}{2}\mu_{2d},} \qquad (5.50)$$

where $E_F(0) = \mu_{2d}$. The result could have been guessed without the math: because of the constant DOS, the average energy is exactly half the maximum energy E_F at $T = 0$ K. This is shown in Fig. 5.10. The energy density per unit area at $T = 0$ K is then given by $u_v(2d) = \frac{1}{2}n_{2d}E_F$. This is the energy cost per unit area to populate the 2D system with non-interacting electrons[18].

At finite temperatures, E_F changes with temperature when the 2D electron gas density is held constant. The d-dimensional quantum

[18] The energy cost increases with the density, meaning if we pack more electrons in the 2D plane, even though they are non-interacting, the exclusion principle makes us pay an energy cost. This is called the *quantum compressibility* of the electron gas.

equipartition of energy $u_d(T) = \frac{d}{2}k_bT\frac{F_{\frac{d}{2}}(\eta)}{F_{\frac{d-2}{2}}(\eta)}$ for $d = 2$ is

$$u_{2d}(T) = k_bT\frac{F_1(\eta)}{F_0(\eta)} \approx \frac{1}{2}\mu_{2d} + s\frac{(k_bT)^2}{\mu_{2d}} \implies c_v^{2d} \approx 2s\frac{k_bT}{\mu_{2d}}n_{2d}k_b, \quad (5.51)$$

where s is a constant of the order unity. The electronic specific heat of the 2D electron system is also lower than the classical value $c_v^{2d}(\text{classical}) = n_{2d}k_b$ by a factor of the order $\frac{k_bT}{\mu_{2d}}$, similar to the case for 1D electron gas, and for exactly the same reason: the Pauli exclusion principle.

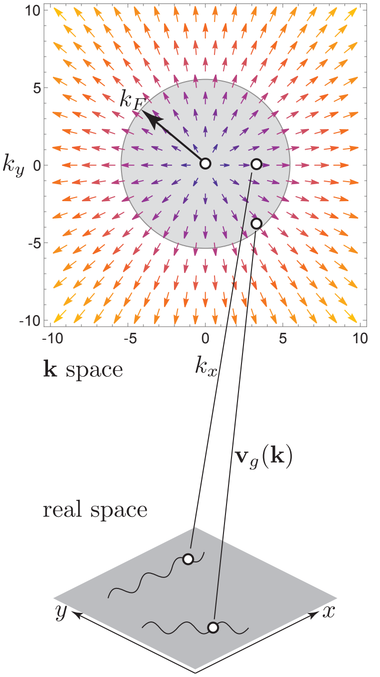


Fig. 5.13 Group velocity of free electrons in 2D in the **k**-space, and its relation to transport in real space.

Quantum Current: We now find the quantum current carried by the 2D electron system. The wavefunction of a state (n_x, n_y) is $\psi(\mathbf{k}, \mathbf{r}) = \frac{1}{\sqrt{A}}e^{i\mathbf{k}\cdot\mathbf{r}}$. The quantum current due to this state is then

$$\mathbf{J}(\mathbf{k}) = \frac{q}{2m_e}(\psi^\star\hat{\mathbf{p}}\psi - \psi\hat{\mathbf{p}}\psi^\star) = q \cdot \frac{1}{A} \cdot \frac{\hbar\mathbf{k}}{m_e} = q(\frac{1}{A})\mathbf{v}_g(\mathbf{k}). \quad (5.52)$$

The group velocity $\mathbf{v}_g(\mathbf{k}) = \frac{\hbar\mathbf{k}}{m_e}$ appears in the expression of the current. The unit of 2D current density is amp/m (or more commonly expressed as mA/μm in semiconductor field-effect transistors), which is a current per unit width. A state $|\mathbf{k}\rangle$ has a group velocity $\mathbf{v}_g(\mathbf{k}) = \frac{\hbar\mathbf{k}}{m_e}$ pointing radially outwards from the origin in the **k**-space, as shown in Fig. 5.13. The corresponding wave velocity in the real space is also shown for two **k**-states, one moving in the $+x$ direction, and another in the $(+x, -y)$ direction. At thermal equilibrium, the occupation of **k**-space is symmetric around the origin. Thus, for every carrier moving radially outward in one direction, there is an exactly equal current in the opposite direction, which means the net current is zero. But the individual currents carried by the **k**-states are substantial!

To find the total current of all occupied electron states, we sum the currents carried by each **k**-state:

$$\boxed{\mathbf{J}_{2d} = q\frac{g_sg_v}{L^2}\sum_{\mathbf{k}}\mathbf{v}_g(\mathbf{k})f(\mathbf{k}) = qg_sg_v\int\frac{d^2k}{(2\pi)^2}\mathbf{v}_g(\mathbf{k})f(\mathbf{k}).} \quad (5.53)$$

Fig. 5.14 (a) shows a schematic of a 2D conductor connected to source and drain contacts. Panel (b) of this figure shows the **k**-space occupation function $f_0(\mathbf{k})$ at equilibrium on the left. In the middle is the distribution function of a ballistic condition when the right-going electron states are in equilibrium with a left contact, and share the same Fermi level E_{Fs} as the left contact, with a corresponding right-going electron Fermi wavevector k_R. The left-going states are in equilibrium with the right-contact, thus sharing the Fermi level E_{Fd} with the drain contact, which fills the left-going states till wavevector k_L. The Fermi circle is split to two halves. The difference of the highest energy states is controlled by the applied voltage. Because the 2D electron density n_{2d} completely determines the area of the Fermi circle at

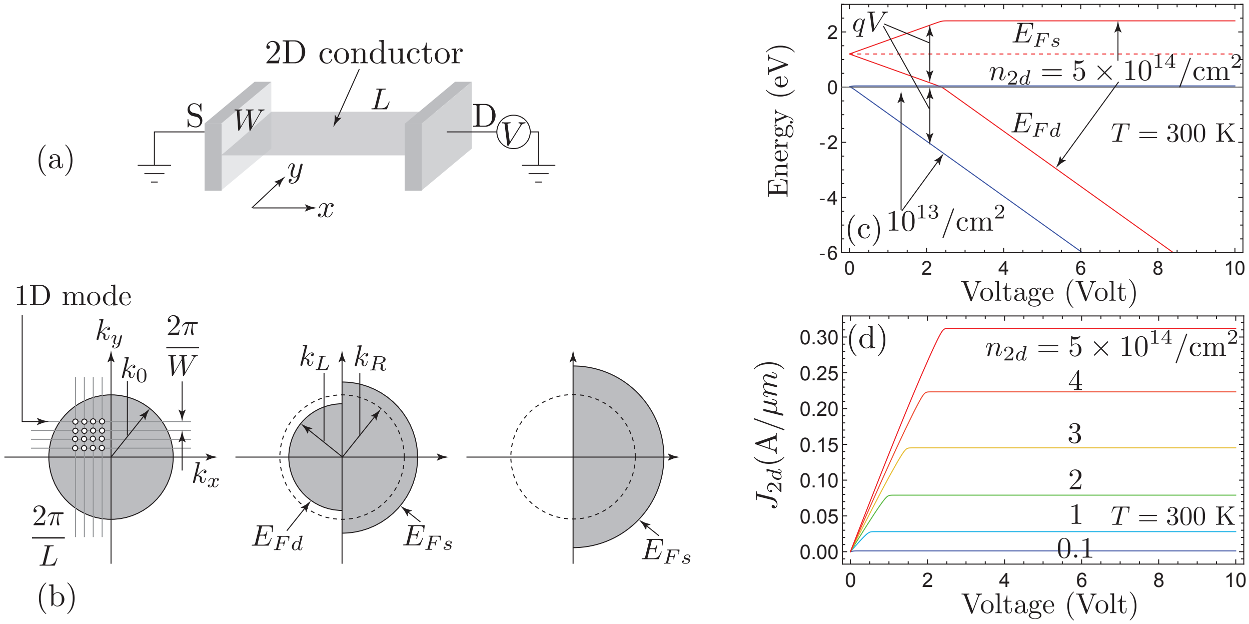


Fig. 5.14 Quantum current flow in 2D electron gases. (a) A 2D conductor is connected to a source and a drain contact, which have much higher number of electron states than the 2D conductor, filled till their respective Fermi levels. The source contact is grounded, and filled to a Fermi level E_{Fs}, and the drain to E_{Fd}, and because of the battery, $E_{Fs} - E_{Fd} = qV$. (b) At $V = 0$, electrons in the 2D conductor fill states (k_x, k_y) up to a well defined Fermi wavevector $k_0 = k_F$, the area of this Fermi circle defined completely by the 2D electron density. When $V \neq 0$, the right going electrons states with $k_x > 0$ are in equilibrium with the source Fermi level E_{Fs}, and the left-going electron states with $k_x < 0$ are in equilibrium with the drain Fermi level E_{Fd}. (c) The calculated split in the quasi Fermi levels E_{Fs} and E_{Fd} for two 2D electron gas densities. The drain voltage at which E_{Fs} saturates is clearly identified around $V = 2$ Volt for $n_{2d} \sim 5 \times 10^{14}/\mathrm{cm}^2$, which is a metallic density, whereas is is very small for $n_{2d} \sim 10^{13}/\mathrm{cm}^3$, closer to a 2D electron gas in a semiconductor nanostructure. (d) The ballistic 2D current for various 2D electron gas densities. The current is expected to saturate, at very high current densities, and at high voltages. This high current density is typically not achieved in 2D metals or semiconductors because of heat dissipation, which is why current saturation is typically not observed in 2D metallic conductors, only the linear portion is observed in experiments. However, saturation is clearly observed when the 2D electron gas density is low, in the $10^{13}/\mathrm{cm}^2$ range.

$V = 0$, if the 2D electron density is fixed, the sum of the areas of the two split halves must remain the same as Fermi circle at $V = 0$. Since the area of the right half circle increases with V, there is a characteristic voltage at which the area of the right-going half circle will become equal to the area of the circle at $V = 0$ as shown in Fig. 5.14 (b). This will mark a crossover into current saturation[19]. Let us now make the above observations quantitative.

If the contacts are ohmic, the total 2D electron density at the source injection point does not change with the applied voltage. Then, we must have

$$n_{2d} = N_c^{2d} \cdot F_0\left(\frac{E_F}{k_b T}\right) = \frac{1}{2} N_c^{2d} \cdot \left[F_0\left(\frac{E_{Fs} - E_c}{k_b T}\right) + F_0\left(\frac{E_{Fd} - E_c}{k_b T}\right)\right], \quad (5.54)$$

which can now be solved with $E_{Fs} - E_{Fd} = qV$ to obtain how the right- and left-going Fermi surfaces change with voltage. The calculated values are shown in Fig. 5.14 (c) for two representative electron densities.

[19] The edge of the filled states define a sharp circle only at $T = 0$ K. At finite temperatures, the edges are diffuse, but still maintain a circular shape.

The behavior is similar to the case discussed earlier for the 1D electron system, and so is the explanation. For a low 2D electron density $n_{2d} = 10^{13}/\text{cm}^2$, E_{Fs} is close to zero, whereas E_{Fd} is qV below E_{Fs}. For a high density $n_{2d} = 5 \times 10^{14}/\text{cm}^2$, E_{Fd} increases with voltage and E_{Fs} decreases. At $V \sim 2.5$ Volt, $E_{Fs} = 0$ and for any further increase in voltage, E_{Fd} clamps to its value at $V \sim 2.5$ Volt and E_{Fs} decreases faster, accounting for all the voltage. This is the condition shown in panel (b) of the figure, when carriers fill the right-going half circle in the **k**-space. The electron group velocities point in various directions making angles $-\frac{\pi}{2} \leq \theta \leq +\frac{\pi}{2}$, but they have $k_x > 0$. Crudely, one can say that most electrons are "moving to the right", and the quantum current should saturate for higher voltages, as the occupation in the **k**-space does not change appreciably.

To find the quantum current flowing to the right, we integrate for all $k_x > 0$ by using the x-component of the group velocity $\mathbf{v_g} \cdot \hat{x} = \frac{\hbar}{m_e} k \cos\theta$, where $k = \sqrt{k_x^2 + k_y^2}$ to obtain

$$J_{2d}^R = \frac{I_d}{W} = \frac{q g_s g_v}{(2\pi)^2} \int_{\theta=-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{k=0}^{\infty} \frac{(\frac{\hbar k \cos\theta}{m_e}) k dk d\theta}{1 + e^{\frac{\frac{\hbar^2 k^2}{2m_e} - E_{Fs}}{k_b T}}} = J_0^{2d} F_{\frac{1}{2}}(\eta_s), \tag{5.55}$$

where

$$J_{2d}^0 = \frac{q^2}{h} \cdot N_c^{1d} \cdot \frac{k_b T}{q}, \tag{5.56}$$

and $N_c^{1d} = g_s g_v (\frac{2\pi m_e k_b T}{h^2})^{\frac{1}{2}}$. The net ballistic current per unit width at a voltage V is then the difference between the right and left going electron currents

$$J_{2d} = J_{2d}^0 \cdot [F_{\frac{1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) - F_{\frac{1}{2}}(\frac{E_{Fs} - E_c - qV}{k_b T})]. \tag{5.57}$$

This current per unit width is plotted in Fig. 5.14 (d) for various 2D electron densities ranging from $n_{2d} = 0.1 - 5 \times 10^{14}/\text{cm}^2$ at $T = 300$ K. Such high saturation currents are typically not observed experimentally in 2D metals because heat dissipation limits the amount of current per unit width to < 10 mA/μm. But in 2D electron gases in semiconductors, the electron densities are in the $n_{2d} \sim 10^{13}/\text{cm}^2$ range, and current densities in the few mA/μm are measured. This is the typical output current per unit width of a field effect transistor that uses a 2D electron gas as the conducting channel.

A second point of observation is unlike the 1D case, the linear or "ohmic" part of the current where $J_{2d} \propto V$ is not the same for all 2D electron gas densities. From Fig. 5.14 (b), the 2D electron system may be considered a set of 1D conductors, each of a fixed k_y. Then it is seen that the applied voltage changes the total number of the 1D modes, and the current may be also evaluated by summing the currents carried by each 1D mode[20]. Using this procedure for evaluating the total current leads to the same result as in Equation 5.57. The

[20] Chapter 17 discusses the mode approach in further detail.

voltage at which the current saturates increases linearly with the 2D electron density as $V_{sat} \propto n_{2d}$.

The saturation current per unit width for the degenerate carrier concentration limit is obtained when $E_{Fs} >> k_b T$ implying $F_{\frac{1}{2}}(\eta_s) \approx \frac{\eta_s^{\frac{3}{2}}}{\Gamma(\frac{5}{2})}$ and $V > V_{sat}$ which implies $F_{\frac{1}{2}}(\eta_d) << F_{\frac{1}{2}}(\eta_s)$. The saturation current per unit width is

$$J_{2d}^{sat} \approx J_{2d}^{0} \frac{\eta_s^{\frac{3}{2}}}{\Gamma(\frac{5}{2})} \approx \frac{q^2}{h} \cdot g_s g_v \left(\frac{2\pi m_e k_b T}{h^2}\right)^{\frac{1}{2}} \cdot \frac{k_b T}{q} \cdot \frac{E_{Fs}^{\frac{3}{2}}}{\frac{3\sqrt{\pi}}{4}(k_b T)^{\frac{3}{2}}}. \tag{5.58}$$

It is seen that the temperature dependence cancels and because $E_{Fs} \propto n_{2d}$, the current per unit width becomes proportional to $n_{2d}^{\frac{3}{2}}$.

In the opposite non-degenerate limit when the 2D electron density is low, when $\eta_s << -1$, and $V > V_{sat}$, the current per unit width is

$$J_{2d}^{sat} \approx J_{2d}^{0} e^{\frac{E_{Fs}-E_c}{k_b T}}, \tag{5.59}$$

which is exponentially dependent on the temperature and the Fermi level.

The two limits encountered above are at the heart of the switching of diodes and transistors; the case when the carrier density is very low results in an exponential dependence of the current on the Fermi level and temperature, and is the **off state**. When the carrier density is high, the system is degenerate with no temperature dependence, and the dependence on the Fermi level is a power law, and is the **on state**.

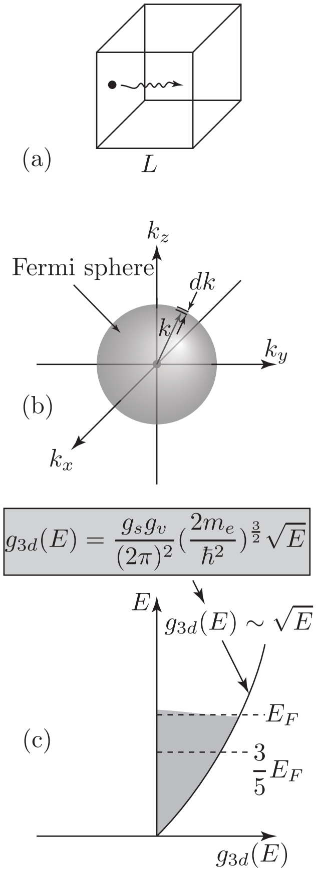

(a)
L
k_z
Fermi sphere
dk
k
k_y
(b)
k_x
$g_{3d}(E) = \frac{g_s g_v}{(2\pi)^2}(\frac{2m_e}{\hbar^2})^{\frac{3}{2}}\sqrt{E}$
E
$g_{3d}(E) \sim \sqrt{E}$
E_F
$\frac{3}{5}E_F$
(c)
$g_{3d}(E)$

Fig. 5.15 (a) An electron in a 3D box of side L can have wavevectors $\mathbf{k} = (k_x, k_y, k_z) = \frac{2\pi}{L}(n_x, n_y, n_z)$ which form a discrete grid in the 3D **k**-space that is shown in (b). When N electrons are filled in the box, the **k**-states are filled inside the Fermi sphere such that the states on the Fermi sphere surface have energy E_F. The density of states for free electrons in 3 dimensions is shown in (c) with a $\sqrt{E}$ dependence, the Fermi level, and the average energy of the electron distribution.

5.5 Electrons in a 3D box

Wavefunctions and Momenta: The electron is now in a cubic box of side L and volume $V = L^3$ as shown in Fig. 5.15 (a) with a spatial coordinate vector $\mathbf{r} = (x, y, z)$, and a wavevector $\mathbf{k} = (k_x, k_y, k_z)$. For $V(\mathbf{r}) = 0$, the allowed wavefunctions are

$$\boxed{\psi(\mathbf{r}) = \frac{1}{\sqrt{L^3}} e^{i(k_x x + k_y y + k_z z)} = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}.} \tag{5.60}$$

The allowed wavevectors are

$$\boxed{\mathbf{k} = (k_{n_x}, k_{n_y}, k_{n_z}) = \frac{2\pi}{L}(n_x, n_y, n_z),} \tag{5.61}$$

and corresponding electron momenta

$$\boxed{\mathbf{p} = \hbar\mathbf{k}, |\mathbf{p}| = \frac{h}{L}\sqrt{n_x^2 + n_y^2 + n_z^2},} \tag{5.62}$$

where n_x, n_y, n_z are again *independent* integers ..., $-2, -1, 0, 1, 2, ...$.

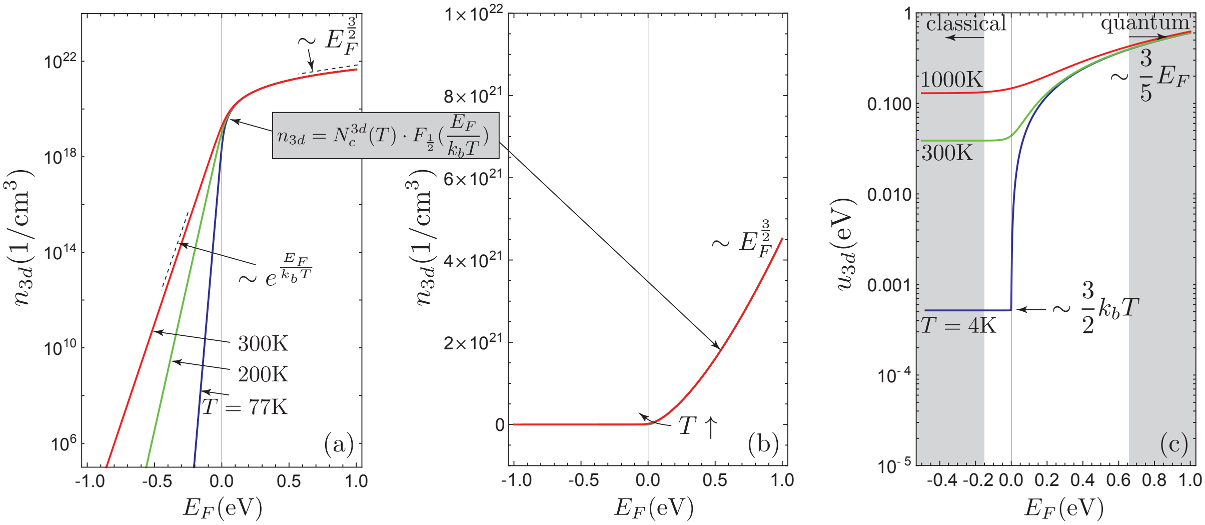


Fig. 5.16 Carrier statistics and energies for the 3D free electron gas. (a) The 3D electron density as a function of the Fermi level E_F governed by $n_{3d} = N_c^{3d} F_{1/2}(E_F/k_b T)$ shown for various temperatures. For high carrier densities (the degenerate limit), the carrier density does not depend on T and depends on E_F as a polynomial power. For low densities (the non-degenerate limit), the dependence on E_F and T is $e^{E_F/k_b T}$, i.e., exponential and very sensitive both to the Fermi level and the temperature. (b) Shows the dependence in the linear scale. It is important to note that though it appears so, $n_{3d} \neq 0$ for $E_F << 0$, but is very small as can be seen in (a). (c) The average energy u_{3d} per electron of the 3D electron gas, which in the non-degenerate limit is the same as the classical equipartition of energy $3 \times \frac{1}{2} k_b T = \frac{3}{2} k_b T$, and in the high density quantum limit is $u_{3d} \sim \frac{3}{5} E_F$ at all temperatures.

Energy and Density of States: In the **k**-space, the allowed set of points now form a 3D grid, each of volume $(\frac{2\pi}{L})^3$. The allowed energy of each such allowed point is

$$\boxed{E(k_x, k_y, k_z) = \frac{\hbar^2}{2m_e}(k_{n_x}^2 + k_{n_y}^2 + k_{n_z}^2) = (n_x^2 + n_y^2 + n_z^2)\frac{h^2}{2m_e L^2} = \frac{\hbar^2 |\mathbf{k}|^2}{2m_e}.} \tag{5.63}$$

The 3D DOS $g_{3d}(E)$ is then

$$g_s g_v \frac{4\pi k^2 dk}{(\frac{2\pi}{L})^3} = G_{3d}(E) dE \implies \frac{G_{3d}(E)}{L^3} = \boxed{g_{3d}(E) = \frac{g_s g_v}{4\pi^2}(\frac{2m_e}{\hbar^2})^{\frac{3}{2}}\sqrt{E}.} \tag{5.64}$$

Here $4\pi k^2 dk$ is the volume of the thin shell of thickness dk around the sphere of radius k in Fig. 5.15 (b). Because each grid point occupies a volume $(\frac{2\pi}{L})^3$ in the **k**-space, there are $\frac{4\pi k^2 dk}{(\frac{2\pi}{L})^3}$ grid points in the shell. Fig. 5.15 (c) shows the resulting density of states.

Because the 3D volume increases as k^3, there are more states at a given energy at higher energies; the $g_{3d}(E) \sim \sqrt{E}$ increase in the 3D DOS is a characteristic feature of 3D electron systems with parabolic dispersion: a result well worth remembering. For the many-electron picture in 3D, we put N non-interacting electrons in the volume V so that the density per unit area is $n_{3d} = N/V$. At $T = 0$ K, the

Pauli exclusion principle suggests states must fill from the center $\mathbf{k} = (0, 0, 0)$ to a sharply defined **Fermi sphere** of radius k_F given by

$$g_s g_v \frac{\frac{4}{3}\pi k_F^3}{(\frac{2\pi}{L})^3} = N \implies \boxed{k_F = (\frac{6\pi^2 n_{3d}}{g_s g_v})^{\frac{1}{3}}.} \qquad (5.65)$$

For $g_s = 2$ and $g_v = 1$, the Fermi wavevector is $k_F = (3\pi^2 n_{3d})^{\frac{1}{3}}$. In a metal with $n_{3d} \sim 10^{24}/\text{cm}^3$, the Fermi wavevector is $k_F \sim 3 \times 10^{10}/\text{m}$, and the Fermi wavelength is $\lambda_F \sim 0.2$ nm, of interatomic dimension. The Fermi surface of the 3D electron gas is the surface of this Fermi sphere. It holds the secrets to most of its properties. For free electrons in 3D, the Fermi surface is spherical. When electrons experience the periodic potential of atoms of a crystal, the Fermi surface can deform and assume a rich range of shapes as we will discuss in later chapters.

The electron density at a temperature T is

$$n_{3d} = \int_0^\infty dE \cdot g_{3d}(E) \cdot f(E) = \frac{g_s g_v}{4\pi^2} (\frac{2m_e}{\hbar^2})^{\frac{3}{2}} \int_0^\infty dE \cdot \sqrt{E} \cdot f(E)$$

$$\implies \boxed{n_{3d} = N_c^{3d} F_{\frac{1}{2}}(\eta),} \qquad (5.66)$$

where the 3D band-edge effective DOS is $N_c^{3d} = g_s g_v (\frac{2\pi m_e k_b T}{h^2})^{\frac{3}{2}}$, and $\eta = E_F/k_b T$. The 3D electron density is shown in Fig. 5.16 (a) and (b) in logarithmic and linear scales as a function of the Fermi energy for various temperatures. The features mimic the properties of 1D and 2D electron systems, characterized by a weak polynomial dependence $n_{3d} \propto E_F^{3/2}$ on the Fermi level at high electron densities and a near independence on temperature in this degenerate limit. In the non-degenerate limit, the density $n_{3d} \approx N_c^{3d} e^{E_F/k_b T}$ is exponentially dependent on both the temperature and the Fermi level.

Part (c) of Fig. 5.16 shows the average energy u_{3d} for various temperatures as a function of the Fermi level. The total energy $\mathcal{U}$ and the average electron energy u at $T = 0$ K is

$$\mathcal{U} = \int_0^\infty dE \cdot E \cdot G_{3d}(E) \cdot f(E)$$

$$\implies \boxed{u_{3d} = \frac{\mathcal{U}}{N} = \frac{\int_0^\infty dE \cdot E \cdot G_{3d}(E) \cdot f(E)}{\int_0^\infty dE \cdot G_{3d}(E) \cdot f(E)} = \frac{3}{5}\mu_{3d},} \qquad (5.67)$$

where $\mu_{3d} = E_F(T = 0K)$. Because the DOS increases with E, the average energy is above $\frac{1}{2}E_F$ at $T = 0$ K. The energy density per unit volume at $T = 0$ K is then given by $\boxed{u_v(3d) = \frac{3}{5} n_{3d} \mu_{3d}}$.

If the 3D electron density is held constant and the temperature increases, the Fermi level must change according to the equation

$$n_{3d} = N_c^{3d} \cdot F_{\frac{1}{2}}(\frac{E_F}{k_b T}) = \frac{1}{3\pi^2}(\frac{2m_e}{\hbar^2}\mu_{3d})^{\frac{3}{2}}. \qquad (5.68)$$

Now using the approximation of the Fermi–Dirac integrals in Table 4.2 for $F_{\frac{1}{2}}(\eta)$, we obtain the result

$$E_F(T) \approx \mu_{3d}[1 - \frac{1}{3}(\frac{\pi}{2}\frac{k_b T}{\mu_{3d}})^2] \tag{5.69}$$

for $k_b T << E_F$. The average energy per electron at a finite temperature is

$$u_{3d}(T) = \frac{3}{2}k_b T \frac{F_{\frac{3}{2}}(\eta)}{F_{\frac{1}{2}}(\eta)} \approx \frac{3}{5}\mu_{3d} + \frac{\pi^2}{4}\frac{(k_b T)^2}{\mu_{3d}}, \tag{5.70}$$

where we have used both the degenerate limit of the Fermi–Dirac integrals, and Equation 5.69, and retained terms to the lowest order in temperature. The average energy density per unit volume in 3D is $u_v(3d) = \frac{3}{5}n_{3d}u_{3d}(T)$, which leads to an electronic contribution to the specific heat

$$c_v^{3d} = \frac{\partial u_v(3d)}{\partial T} \approx \frac{\pi^2}{2} \cdot \frac{k_b T}{\mu_{3d}} \cdot n_{3d}k_b. \tag{5.71}$$

Fig. 5.17 Arnold Sommerfeld first introduced elliptical orbits to Bohr's model of the atom. He introduced the quantum electron theory of metals by changing the electron statistics from the classical thermodynamic relations to the then-newly discovered Fermi–Dirac distribution after learning from Pauli of his exclusion principle. Doing so resolved some of the major discrepancies of the Drude model such as the much lower electronic specific heat. The Sommerfeld model neglected the periodic potential of crystals experienced by electrons, it was essentially a free electron model that includes scattering. Sommerfeld advised and mentored of a large cohort of scientists and several Nobel Prize winners, but was never awarded the prize himself in spite of being nominated ~80 times!

This result was first obtained famously by Sommerfeld (Fig. 5.17) using (what is now called) the Sommerfeld expansion of Fermi–Dirac integrals. It showed for the first time why the electronic contribution to the specific heat in a 3D metal is much lower than the classical thermodynamic limit $\frac{3}{2}n_{3d}k_b$. The reason is the same as discussed for the 1D and 2D cases: only a fraction $\frac{k_b T}{\mu_{3d}}$ of all the electrons are able to absorb $k_b T$ of energy upon heating because these electrons are at the Fermi level: for electrons with energy much lower than the Fermi levels, the Pauli exclusion principle forbids them from absorbing heat and increasing their energies.

Quantum Current: The quantum current carried by the 3DEG is obtained by summing the contributions of all occupied **k**-states:

$$\boxed{\mathbf{J}_{3d} = q\frac{g_s g_v}{L^3}\sum_{\mathbf{k}} \mathbf{v}_g(\mathbf{k})f(\mathbf{k}) = qg_s g_v \int \frac{d^3k}{(2\pi)^3}\mathbf{v}_g(\mathbf{k})f(\mathbf{k}),} \tag{5.72}$$

which has units of Amp/m^2, or current per unit area. Similar to the 2D case, a state $|\mathbf{k}\rangle$ has a group velocity $\mathbf{v}_g(\mathbf{k}) = \frac{\hbar\mathbf{k}}{m_e}$ pointing radially outwards from the origin of the 3D **k**-space shown in Fig. 5.18 (b). The net current at thermal equilibrium is zero because for every $|+\mathbf{k}\rangle$ state, there is a corresponding $|-\mathbf{k}\rangle$ state. This delicate balance is broken when a voltage is applied on the drain. The voltage does not change the total electron density at the source injection point, which means the condition

$$n_{3d} = \frac{1}{2}N_c^{3d} \cdot [F_{\frac{1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) + F_{\frac{1}{2}}(\frac{E_{Fs} - E_c - qV}{k_b T})] \tag{5.73}$$

should be satisfied. Solving this condition with $E_{Fs} - E_{Fd} = qV$ allows us to obtain the E_{Fs} for a given n_{3d}. A few calculated values are shown in Fig. 5.18 (c). To find the current flowing in the x-direction, we must

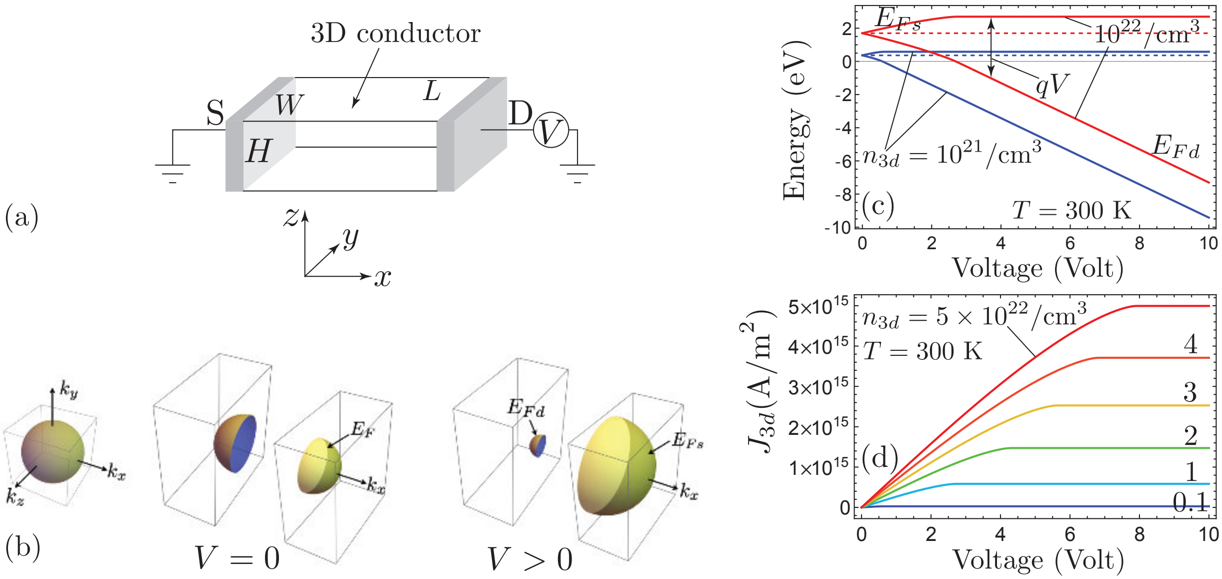


Fig. 5.18 Quantum current flow in 3D electron gases. (a) A 3D conductor is connected to a source and a drain contact and a voltage V is applied. (b) At $V = 0$, electrons in the 3D conductor fill states (k_x, k_y, k_z) up to a Fermi wavevector $k_0 = k_F$. The volume of the Fermi sphere is defined completely by the 3D electron density n_{3d}. When $V \neq 0$, the right going electrons states in the hemisphere with $k_x > 0$ are in equilibrium with the source Fermi level E_{Fs}, and the left-going electron states in the hemisphere $k_x < 0$ are in equilibrium with the drain Fermi level E_{Fd}. The hemispheres are shown split for clarity: the effect of the voltage is to increase the right going hemisphere volume and decrease the left-going volume while maintaining the same total volume. (c) The calculated split in the quasi Fermi levels E_{Fs} and E_{Fd} for two 3D electron gas densities. The drain voltage at which E_{Fs} saturates is is higher for higher n_{3d}. (d) The ballistic 3D current density in A/cm^2 for various 3D electron gas densities. The current saturates at very high current densities, and at high voltages. In the linear part of the current-voltage curve, the current depends even more strongly on n_{3d} than it did in n_{2d}, and saturation sets in at voltages that depend on the 3D electron density as $V_{sat} \propto n_{3d}^{2/3}$.

first take the projection of the group velocities of all filled **k**-states along the x-axis, i.e., $v_g = \mathbf{v}_g(\mathbf{k}) \cdot \hat{x}$ and sum this over all states. It is convenient here to imagine the k_x-axis to be the z-coordinate in a spherical system (k, θ, ϕ). Upon doing so, the integral for the current from Equation 5.72 for the right-going states reduces to

$$J_{3d}^R = q g_s g_v \int_{k=0}^{\infty} \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\frac{\pi}{2}} \frac{k^2 \sin\theta d\theta d\phi dk}{(2\pi)^3} \cdot \frac{\frac{\hbar k}{m_e}\cos\theta}{1 + e^{\frac{E_c + \frac{\hbar^2 k^2}{2m_e} - E_{Fs}}{k_b T}}}$$

$$\implies J_{3d}^R = \frac{q^2}{h} \cdot N_c^{2d} \cdot \frac{k_b T}{q} \cdot F_1(\eta_s), \quad (5.74)$$

and the total current density is then

$$\boxed{J_{3d} = \frac{q^2}{h} \cdot N_c^{2d} \cdot \frac{k_b T}{q} \cdot [F_1(\frac{E_{Fs} - E_c}{k_b T}) - F_1(\frac{E_{Fs} - E_c - qV}{k_b T})].} \quad (5.75)$$

Here $N_c^{2d} = g_s g_v (\frac{2\pi m_e k_b T}{h^2})$, the band-edge DOS of 2D. This result is consistent with the d-dimensional general result that the quantum current in d-dimensions is $J_d = \frac{q^2}{h} \cdot N_c^{d-1} \cdot \frac{k_b T}{q} \cdot [F_{\frac{d-1}{2}}(\eta_s) - F_{\frac{d-1}{2}}(\eta_d)]$, with $d = 3$. Fig. 5.18 (d) shows the calculated quantum current density in A/cm^2 for a few 3D electron densities. The features of an Ohm's law like increase at small voltages, followed by current saturation is similar to the cases of transport in the 1D and 2D electron systems. The linear regime shows a stronger dependence of the current on the electron density than the 2D case.

Electrons have a range of velocities in the quantum mechanical case ranging from zero, to along the current direction and also opposite to the current direction. The simplest classical picture of current flow is the following: if there are n_d electrons per unit volume, the current is $J_d = qn_d \langle v \rangle_d$, where it is assumed that *all* electrons are moving in the direction of the current flow with the same effective ensemble velocity $\langle v \rangle_d$. Though it is not microscopically correct, the ensemble velocity is a measure of the conductivity of the electron system. If we equate this classical picture with the quantum version,

$$J_d = qn_d \langle v \rangle_d = \frac{q^2}{h} \cdot N_c^{d-1} \cdot \frac{k_b T}{q} \cdot [F_{\frac{d-1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) - F_{\frac{d-1}{2}}(\frac{E_{Fs} - E_c - qV}{k_b T})], \tag{5.76}$$

and using $n_d = N_c^d F_{\frac{d-2}{2}}(\eta_F)$, we can obtain an expression for the effective classical ensemble electron velocity in d-dimensions:

$$\langle v \rangle_d = \sqrt{\frac{k_b T}{2\pi m_e}} \cdot \frac{F_{\frac{d-1}{2}}(\eta_s) - F_{\frac{d-1}{2}}(\eta_d)}{F_{\frac{d-2}{2}}(\eta_F)} \approx \sqrt{\frac{E_F}{2\pi m_e}} \cdot 2^{\frac{d+1}{d}} \cdot \frac{\Gamma(\frac{d+2}{2})}{\Gamma(\frac{d+3}{2})}, \tag{5.77}$$

where the approximation on the right holds for the highly degenerate electron distribution. The effective ensemble electron velocities for three temperatures, for all three dimensions are shown in Fig. 5.19. Under degenerate conditions, the ensemble velocity is the highest for 1D electron systems, which should not come as a surprise because all electrons are moving along the direction of the flow of current, unlike in the 2D and 3D cases. From the degenerate approximation in the equation above, the temperature cancels, and the effective ensemble velocity increases as $\sqrt{E_F}$. In the non-degenerate limit, the temperature has the strongest effect on the ensemble velocities, and the effect of dimensionality vanishes because the factor of the Fermi–Dirac integrals ratio becomes unity in the Maxwell–Boltzmann limit, leaving an identical effective ensemble velocity of $\sqrt{k_b T / 2\pi m_e}$ for all dimensions.

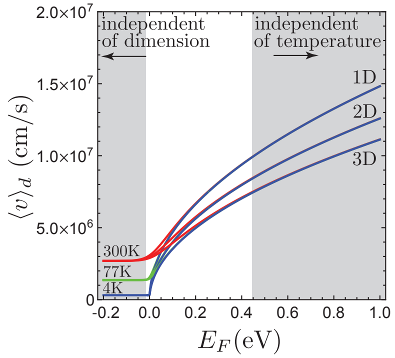


Fig. 5.19 Effective ensemble velocities for free electron transport in 1D, 2D, and 3D at three characteristic temperatures at a voltage of $V = 5$ Volts.

This concludes our discussion of the quantum properties of free electrons in various dimensions. The various properties that were discussed are listed in a succinct form in Table 5.1 for easy reference. Much of the properties of semiconductors and nanostructures that we will encounter in later chapters will require some of these relations, but with modified effective masses because of the appearance of bands

Table 5.1 Quantum properties of free electron dases in different dimensions

Property	1D	2D	3D
k_n	$\frac{2\pi}{L_x}n_x$	$(\frac{2\pi}{L_x}n_x, \frac{2\pi}{L_x}n_y)$	$(\frac{2\pi}{L_x}n_x, \frac{2\pi}{L_y}n_y, \frac{2\pi}{L_z}n_z)$
ψ_n	$\frac{e^{ik_x x}}{\sqrt{L_x}}$	$\frac{e^{ik_x x}}{\sqrt{L_x}} \cdot \frac{e^{ik_y y}}{\sqrt{L_y}}$	$\frac{e^{ik_x x}}{\sqrt{L_x}} \cdot \frac{e^{ik_y y}}{\sqrt{L_y}} \cdot \frac{e^{ik_z z}}{\sqrt{L_z}}$
p_n	$\hbar(\frac{2\pi}{L_x}n_x)$	$\hbar(\frac{2\pi}{L_x}n_x, \frac{2\pi}{L_y}n_y)$	$\hbar(\frac{2\pi}{L_x}n_x, \frac{2\pi}{L_y}n_y, \frac{2\pi}{L_z}n_z)$
$E_n(k)$	$\frac{\hbar^2}{2m_e}k_x^2$	$\frac{\hbar^2}{2m_e}(k_x^2+k_y^2)$	$\frac{\hbar^2}{2m_e}(k_x^2+k_y^2+k_z^2)$
$k_F(0K)$	$\frac{\pi n_{1d}}{g_s g_v}$	$(\frac{4\pi n_{2d}}{g_s g_v})^{\frac{1}{2}}$	$(\frac{6\pi^2 n_{3d}}{g_s g_v})^{\frac{1}{3}}$
$g_d(E)$	$\frac{g_s g_v}{2\pi}(\frac{2m_e}{\hbar^2})^{\frac{1}{2}}\frac{1}{\sqrt{E}}$	$\frac{g_s g_v m_e}{2\pi\hbar^2}\Theta(E)$	$\frac{g_s g_v}{4\pi^2}(\frac{2m_e}{\hbar^2})^{\frac{3}{2}}\sqrt{E}$
$u_v(0K)$	$\frac{1}{3}n_{1d}E_F$	$\frac{1}{2}n_{2d}E_F$	$\frac{3}{5}n_{3d}E_F$
$E_F^d(T)$	$E_F^{1d}(0)[1+\frac{\pi^2}{12}(\frac{k_bT}{E_F^{1d}(0)})^{\frac{5}{2}}]$	$k_bT\ln(e^{\frac{E_F^{2d}(0)}{k_bT}}-1)$	$E_F^{3d}(0)[1-\frac{\pi^2}{12}(\frac{k_bT}{E_F^{3d}(0)})^2]$
$u_v^d(T)$	$\frac{1}{3}E_F^{1d}(0)+\frac{2\pi^2}{9}\frac{(k_bT)^2}{E_F^{1d}(0)}$	$u_v^{1d}(0K)(1-\frac{k_bT}{E_F})$	$\frac{3}{5}E_F^{3d}(0)+\frac{\pi^2}{4}\frac{(k_bT)^2}{E_F^{3d}(0)}$
$c_v^d(T)$	$\frac{4\pi^2}{9}\cdot\frac{k_bT}{E_F^{1d}(0)}\cdot n_{1d}k_b$	$\frac{\pi^2}{2}(\frac{k_bT}{E_F^{3d}(0)})n_{3d}k_b$	$\frac{\pi^2}{2}\cdot\frac{k_bT}{E_F^{3d}(0)}\cdot n_{3d}k_b$
$\frac{n_d(T)}{g_d(k_bT)}$	$k_bT\cdot\sqrt{\pi}F_{-\frac{1}{2}}(\frac{E_F}{k_bT})$	$k_bT\cdot F_0(\frac{E_F}{k_bT})$	$k_bT\cdot\sqrt{\frac{\pi}{2}}F_{+\frac{1}{2}}(\frac{E_F}{k_bT})$
$n_d(T)$	$N_c^{1d}F_{-\frac{1}{2}}(\frac{E_F^{1d}(T)}{k_bT})$	$N_c^{2d}F_0(\frac{E_F^{2d}(T)}{k_bT})$	$N_c^{3d}F_{+\frac{1}{2}}(\frac{E_F^{3d}(T)}{k_bT})$
J_R	$\frac{qg_sg_v}{2\pi\hbar}(k_bT)F_0(\frac{E_{Fs}}{k_bT})$	$\frac{qg_sg_v}{(2\pi)^{\frac{3}{2}}}\frac{\sqrt{m_e}}{\hbar^2}(k_bT)^{\frac{3}{2}}F_{\frac{1}{2}}(\frac{E_{Fs}}{k_bT})$	$\frac{qg_sg_v}{(2\pi)^2}\frac{m_e}{\hbar^3}(k_bT)^2F_1(\frac{E_{Fs}}{k_bT})$

in periodic solids, instead of the free electron parabolic bandstructure discussed in this chapter.

For the free electron, the electric potential seen by the electron is $V(\mathbf{r}) = 0$. Before we leave this chapter, we consider a few situations when the electron is not free, but bound to atoms or molecules. These situations are modeled in quantum mechanics with simple potentials $V(\mathbf{r})$. We discuss a few potentials that will appear in later chapters when we consider quantum heterostructures, dopants, and the interaction of electrons with photons[21].

5.6 The particle in a box

The electron is placed in a 1D region that has a potential $V(x) = 0$ for $0 < x < L$, and $V(x) = \infty$ outside, as shown in Fig. 5.20. Applying the same tricks that we have done earlier in this chapter for the free electron, we first solve the Schrödinger equation

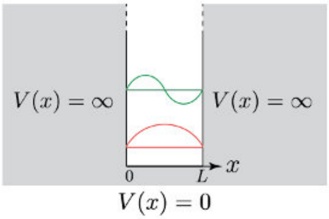
$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x) + V(x)\psi(x) = E\psi(x), \tag{5.78}$$

for the electron with the potential

$$V(x) = 0, \quad 0 \le x \le L \tag{5.79}$$

$$V(x) = \infty, \quad x < 0, x > L. \tag{5.80}$$

[21] The discussion in the following sections is kept terse since the material is covered in introductory textbooks on quantum mechanics.



Particle in a box

Fig. 5.20 Electron in a 1D box.

The major difference from the free electron case is that $\psi(x) = 0$ in regions where $V(x) = \infty$. We still try plane wave solutions, but with the boundary condition that the wavefunction goes to *zero* at the edges of the box:

$$\psi(x) = Ae^{ikx} + Be^{-ikx} \rightarrow \psi(0) = 0 = A + B, \psi(L) = Ae^{ikL} + Be^{-ikL} = 0 \tag{5.81}$$

which implies

$$\frac{A}{B} = -e^{-i2kL} = -1 \rightarrow 2kL = 2n\pi \rightarrow \boxed{k_n = n\frac{\pi}{L},} \; n = \pm1, \pm2, \pm3, ... \tag{5.82}$$

Note that $n = 0$ is *not allowed*, because then $\psi(x) = 0$ and there is no particle. The wavefunction after normalization over the length L is

$$\psi_n(x) = \sqrt{\frac{2}{L}} \sin(n\frac{\pi}{L}x) = \sqrt{\frac{2}{L}} \sin(k_n x), \tag{5.83}$$

and the allowed energy eigenvalues are

$$E_n = n^2 \frac{(\pi\hbar)^2}{2m_e L^2} = n^2 \frac{h^2}{8m_e L^2}. \tag{5.84}$$

The results look similar to the 1D free electron problem, but this appearance is deceptive. The eigenfunctions $\psi_n(x)$ are only valid between $0 < x < L$, meaning the wavefunction at $x = 2L$ is not $\psi_n(x = 2L)$, it is zero. What this means is that since the electron is confined in real space to $\Delta x = L$, the momentum of the states must be spread out by $\Delta k_x \geq \frac{1}{2L}$. *Hence the energy eigenstates do not have a definite momentum.* Each energy eigenstate has a range of momenta, centered around zero. The expectation value of the momentum for all states is identically zero. The physical meaning of this result is seen most clearly if we calculate the quantum current of any eigenstate: because the eigenfunction is real, the quantum current $J = \frac{-i\hbar q}{2m_e}(\psi\partial_x\psi^\star - \psi^\star\partial_x\psi) = 0$ is identically zero. This means that electrons that occupy any eigenstate of the quantum well do not carry any current – they are really not going anywhere outside the quantum well, they are not free to move, but **bound** inside the well[22].

The states in the well follow the Pauli exclusion principle. Each state has a spin degeneracy of $g_s = 2$. For example, if there are three electrons in the well, the lowest energy state of the system is when two electrons with opposite spins fill eigenvalue E_1, and the last electron with any spin occupies eigenvalue E_2, making the total energy of the system $2E_1 + E_2$. Because the energy states are discrete, and sharp, the analog of the density of states defined earlier for free electrons is a series of Dirac delta functions $g(E) = \sum_n \delta(E - E_n)$. The particle in the box in higher dimensions in 2D and 3D has simple results just like the 1D case, but with additional quantum numbers n_y and n_z, with the energies still discrete and sharp, but some additional degeneracies of energy eigenvalues set in due to the choices quantum numbers.

[22] A semi-classical picture that is sometimes useful is to imagine the bound states are as standing waves formed due to the interference of left- and right-going electron waves. This is analogous to light trapped in an optical cavity bouncing between two mirrors.

If the barrier height is not infinite, then we have a finite number of bound states that have energies smaller than the barrier, and the energies higher than the barrier height are not bound. Furthermore, the finite barrier height allows the wavefunctions of the bound states to have a finite penetration into the barrier.

The particle in the box problem of quantum mechanics is the simplest, and most powerful model for any bound state problem. Though there are no 1D atoms, it is possible to capture most results of atomic physics by imagining the complicated potentials that atoms can exert on electrons by the simple finite, or infinite barrier height particle in a box.

5.7 The Dirac delta potential

Consider the following potential seen by electrons (Fig. 5.21):

$$V(x) = -V_0, \quad -\tfrac{L}{2} \leq x \leq \tfrac{L}{2} \tag{5.85}$$

$$V(x) = 0, \quad x < -\tfrac{L}{2}, x > \tfrac{L}{2}. \tag{5.86}$$

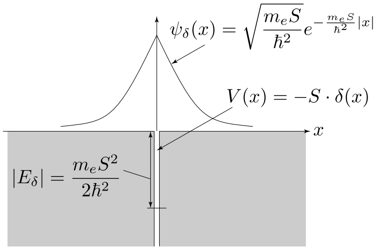

$\psi_\delta(x) = \sqrt{\frac{m_e S}{\hbar^2}} e^{-\frac{m_e S}{\hbar^2}|x|}$
$V(x) = -S \cdot \delta(x)$
x
$|E_\delta| = \frac{m_e S^2}{2\hbar^2}$

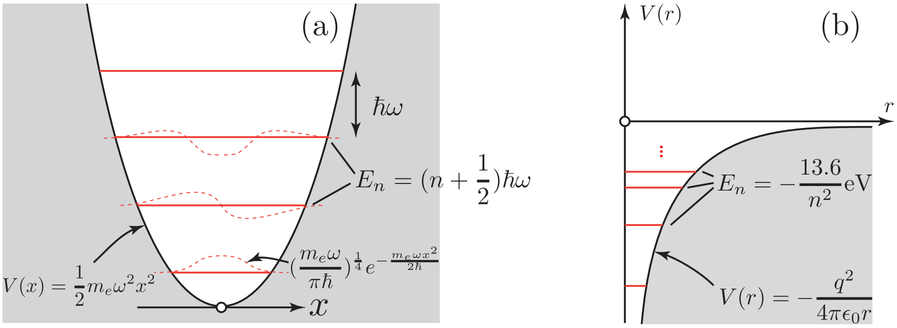
Fig. 5.21 The 1D Dirac delta well.

This a trap for the electron that extends over a length L and is V_0 units deep in energy. Let us decrease the length $L \downarrow$ and make the well deeper $V_0 \uparrow$, while keeping the product or "strength" of the trap $S = V_0 L$ constant. In the limit $L \to 0$, $V_0 \to \infty$, meaning the potential must become infinitely deep to keep S the same. This limit is the Dirac delta potential, written simply as $V(x) = -S\delta(x)$. For the 1D Dirac delta potential, there is **only one bound state** with energy eigenvalue and eigenfunction

$$\boxed{E_\delta = -\frac{m_e S^2}{2\hbar^2},} \text{ and } \boxed{\psi_\delta(x) = \sqrt{\frac{m_e S}{\hbar^2}} e^{-\frac{m_e S}{\hbar^2}|x|}.} \tag{5.87}$$

There are of course lots of propagating states with $E > 0$ that are scattered by the Dirac delta potential well. An important characteristic of the **1D Dirac delta potential well is that it admits exactly one bound state**. One may naively think that if the strength of the potential well S is very shallow either due to V_0 being small, or L being very narrow, the bound state is pushed out of the well. Or that there are more than one bound states for a stronger binding strength. But this is not correct in 1D – there is exactly one bound state *no matter how weak or strong the binding strength.*

The Dirac delta potential well in 2D also allows exactly one bound state. However, unlike in 1D, the 2D bound state becomes exponentially shallow as the binding strength decreases to very small values. In a 3D Dirac delta potential well, if the binding strength is very weak, there may be no bound states at all.


(a)
$\hbar\omega$
$E_n = (n + \frac{1}{2})\hbar\omega$
$(\frac{m_e\omega}{\pi\hbar})^{\frac{1}{4}} e^{-\frac{m_e\omega x^2}{2\hbar}}$
$V(x) = \frac{1}{2} m_e \omega^2 x^2$
x
$V(r)$
(b)
r
$E_n = -\frac{13.6}{n^2}$eV
$V(r) = -\frac{q^2}{4\pi\epsilon_0 r}$

Harmonic Oscillator

Hydrogen Atom

Fig. 5.22 (a) The harmonic oscillator potential, and (b) the radially attractive Coulomb potential of a hydrogen atom. The energy eigenvalues for each case are indicated.

5.8 The harmonic oscillator

Fig. 5.22 (a) shows a parabolic potential of a 1D harmonic oscillator

$$V(x) = \frac{1}{2} m_e \omega^2 x^2. \tag{5.88}$$

Solving the Schrödinger equation yields that the allowed electron energy eigenvalues for this potential are *equally spaced* in energy:

$$E_n = (n + \frac{1}{2})\hbar\omega, \tag{5.89}$$

where ω is a measure of the curvature of the parabolic potential, and $n = 0, 1, 2, ...$ The ground state energy $\frac{1}{2}\hbar\omega$ like the infinite square well potential is nonzero. The spacing of energies is exactly $\hbar\omega$. Unlike the infinite square well potential, in which the eigenvalues increase as $E_n \sim n^2$, they increase *linearly* as $E_n \sim n$ for the oscillator[23]. The wavefunctions are Hermite polynomials as indicated in Fig. 5.22 (a).

5.9 The hydrogen atom

The hydrogen atom has one proton and one electron. Fig. 5.22 (b) shows that the Coulomb potential energy of an electron at a distance r from the proton is

$$V(r) = -\frac{q^2}{4\pi\epsilon_0 r}. \tag{5.90}$$

The 3D Schrödinger equation for the electron in this potential

$$-\frac{\hbar^2}{2m_e}(\partial_x^2 + \partial_y^2 + \partial_z^2)\psi(x,y,z) + V(r)\psi(x,y,z) = E\psi(x,y,z) \tag{5.91}$$

is solved in the spherical coordinate system by exploiting the symmetry of the potential. This was done famously for the first time by

---

[23] The uniform spacing is a very good model for phonons and photons, which are wave phenomena, as will be described in later chapters.

Schrödinger with help from a mathematician, and gained his equation instant popularity and fame. The energy eigenvalues are

$$E_n = -\frac{m_e q^4}{2(4\pi\epsilon_0)^2\hbar^2}\frac{1}{n^2} = -\frac{13.6}{n^2}\ \text{eV}, \tag{5.92}$$

with corresponding eigenfunctions $\psi(r,\theta,\phi) = R(r)\Theta(\theta)\Phi(\phi)$ that have four quantum numbers: (n,m,l,s) which capture the radial part, the angular momentum of electrons around the proton, and the spin of the electron. The eigenvalues above are shown in Fig. 5.22 (b). They are exactly the same as Bohr had obtained from this orbit theory as was discussed in Chapter 3. But Schrödinger's equation for the first time gave a firm mathematical basis for the spectrum of the hydrogen atom, and opened up the path for understanding complicated molecules and solids.

The few bound state problems discussed above have exact analytical solutions of the electron eigenvalues and eigenfunctions. There are few more potentials for which exact solutions may be obtained, such as a half-parabolic harmonic oscillator which allows only the even eigenvalues of the full harmonic oscillator, or for which approximate analytical solutions may be obtained, such as the triangular quantum well which leads to Airy function solutions. With the solutions to the bound states of the electron for these simple potentials, we are well armed to handle the several classes of problems that we will encounter in the later chapters in this book. If more complicated potentials are encountered, the Schrödinger equation can be solved approximately by first finding the exactly solved potential problem that is closest, and then applying perturbation theory, which is the subject of the next two chapters.

5.10 Chapter summary section

In this chapter, we learned:

- The momentum, eigenfunction, eigenvalue spectra of electrons free to move in one, two, and three spatial dimensions,
- The thermodynamic properties such as average energy and specific heat of free electrons in one, two, and three dimensions, and how the fact that the electronic part of the specific heat is lowered significantly by enforcing the Pauli exclusion principle,
- The quantum mechanical current flowing due to free electrons in 1D, 2D, and 3D conductors upon the application of a voltage across contact electrodes. The current exhibits a characteristic linear regime at small voltages that looks like Ohm's law but differs in several ways from the classical picture, and current saturation at high voltages that has no classical analog, and

- The eigenvalues and eigenfunctions of a few exactly solvable bound-state problems for the electron.

Further reading

The free electron, and various bound state problems are covered well in various textbooks of quantum mechanics, of which **Introduction to Quantum Mechanics** by D. J. Griffiths, and **Quantum Mechanics** by C. Cohen-Tannoudji and B. Diu are recommended. The quantum mechanics and quantum thermodynamics of the 3D free electron problem is discussed in significant detail and with insight in the introductory chapters of **Solid State Physics** by N. Ashcroft and D. Mermin. The quantum transport problem is presented under the umbrella of semiconductor nanostructures and devices in **Nanoscale Transistors** by M. Lundstrom and J. Guo.

Exercises

(5.1) **Born to be free**

As discussed in this chapter, because of the Pauli exclusion principle, fermions must follow the Fermi–Dirac distribution, and they have half-integer spins. Now imagine we have a metal with $n = 10^{23}/\text{cm}^3$ electrons in a cubic box of side L, and we know that electrons are fermions. Assume the electrons are completely free to move around in the box, meaning there are no atoms in their way. If that that much freedom is not enough for you, how about this: completely neglect the Coulomb interactions due the charge of the electrons! (This may all be very unsettling, but we will explain later why it is actually OK to do so – because with great freedom comes great responsibility! In fact this problem could be formulated for *any* fermion – for example the uncharged neutron – and the analytical answers will be the same.) Find the following at $T = 0\text{K}$:

(a) The Fermi wavevector k_F.

(b) The Fermi momentum p_F.

(c) The Fermi energy E_F.

(d) The *average* energy of electrons $u = \frac{U}{N}$. What is the origin of this energy?

(e) What is the average energy of the electrons if they did not follow quantum mechanics, but were subject to classical mechanics?

(5.2) **Graphene DOS and Fermi–Dirac distribution**

The electrons in the conduction band of graphene are free to move in 2-dimensions, forming a 2DEG. The energy-momentum dispersion relationship for the 2DEG electrons in graphene is $E(k_x, k_y) = \hbar v_F \sqrt{k_x^2 + k_y^2}$, where v_F is a parameter with dimensions of velocity. For graphene, it is $v_F = 10^8 \text{cm/s}$.

(a) Make a sketch of the energy as a function of the (k_x, k_y) points in the 2D k-space plane, and show that the dispersion results in a conical shape.

(b) Show that the density of states for these electrons is $g(E) = \frac{g_s g_v}{2\pi(\hbar v_F)^2}|E|$, where $g_s = 2$ is the spin degeneracy of each (k_x, k_y) state, and g_v is the number of cones in the energy dispersion. For graphene, $g_v = 2$.

(c) Show that at thermal equilibrium, when the Fermi level is at $E_f = 0$, the number of conduction electrons per unit area in 2D graphene is $n_i = \frac{\pi}{6}(\frac{kT}{\hbar v_F})^2$. Make a plot of this density as a function of temperature for $0K \leq T \leq 500K$. Explain why your plot sets the bar on the lowest possible density of carriers achievable in graphene at those temperatures.

Geim
Novoselov

Fig. 5.23 Andre Geim and Kostya Novoselov were awarded the Nobel Prize in physics in 2010 for the discovery of graphene, the thinnest 2D crystal with remarkable electron transport properties.

(5.3) Density of states of electrons, photons, and phonons

(a) Show that for a parabolic bandstructure for electrons $E(k) = E_c + \frac{\hbar^2 k^2}{2m^\star}$ with band edge E_c and effective mass $m^\star$, the DOS for electron motion in d dimensions is

$$\boxed{g_d(E) = \frac{g_s g_v}{2^d \pi^{\frac{d}{2}} \Gamma(\frac{d}{2})} (\frac{2m^\star}{\hbar^2})^{\frac{d}{2}} (E - E_c)^{\frac{d}{2}-1},} \tag{5.93}$$

where g_s is the spin degeneracy, and g_v is the valley degeneracy. Here $\Gamma(...)$ is the Gamma function with property $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. You may need the expression for the surface area of a d-dimensional sphere in k-space: $S_d = \frac{2\pi^{\frac{d}{2}} k^{d-1}}{\Gamma(\frac{d}{2})}$. Check that this reduces to the surface area of a sphere for $d = 3$ and the circumference of a circle for $d = 2$.

(b) Sketch the DOS for 3D, 2D, and 1D electron systems using the expression. Explain the roles of the valley degeneracy and the effective mass for silicon and compound semiconductors.

(c) Show that the DOS for energy dispersion $E(k) = \hbar v k$ for three dimensions is

$$\boxed{g_\omega(\omega) = \frac{g_p \omega^2}{2\pi^2 \hbar v^3},} \tag{5.94}$$

where $\omega = vk$, and g_p is the polarization degeneracy. This is the dispersion for *waves*, such as photons and phonons moving with velocity v. The parabolic DOS of phonons and photons will play an important role in the thermal and photonic properties of semiconductors.

(5.4) Periodic vs. hard-wall boundary conditions

Instead of the periodic boundary conditions that require that the wavefunction at the "ends" of a region in space just be equal (e.g. $\psi(x) = \psi(x+L)$), one may enforce what are called hard-wall boundary conditions for the 1D, 2D, or 3D electron gas. The hard-wall boundary condition requires that the wavefunction vanish at the edges, for example for the 1D electron gas at $x = 0$ and at $x = L$. Show that the allowed k-states are then rearranged: instead of the spacing for periodic boundary condition $k_n = \frac{2\pi}{L}n$ where $n = 0, \pm 1, \pm 2, ...$, the hard wall states are spaced closer: $k_n = \frac{\pi}{L}n$ where $n = 1, 2, 3, ...$ Show that though the k-space spacing changes, the density of states remains the same. Show that for the 2D case, instead of $\mathbf{k}$-states allowed for the hard-wall boundary condition occupies only the first quadrant of the $\mathbf{k}$-space, but again with the same DOS. Similarly, show that for the 3D free electron gas, only the first octant of the 3D $\mathbf{k}$-space hosts the allowed $\mathbf{k}$-states, and the DOS is the same as that worked out for the periodic boundary condition.

(5.5) Sommerfeld's coup

(a) Using the DOS you calculated in Exercise 5.3, find the *total energy* of N electrons in volume V at $T = 0$ K for 3D, 2D, and 1D electron gases with parabolic energy dispersion. Note that you already solved the 3D electron gas problem in Exercise 5.1.

(b) Now for the heat capacity $c_v = \frac{1}{V}\frac{dU}{dT}$, we need to find the total energy U at a non-zero temperature T. To do that, you can still use the fact that heating a bunch of electrons will not increase or decrease their number. Show (this is somewhat hard, but Sommerfeld did it ~100 years ago!) that for 3D electrons, the Fermi energy changes with temperature as

$$E_F(T) = E_F(0)[1 - \frac{1}{3}(\frac{\pi k_b T}{2E_F(0)})^2]. \tag{5.95}$$

(c) Show that the heat capacity of 3D "quantum" electrons is then

$$c_v = \frac{\pi^2}{2} n k_b \frac{k_b T}{E_F(0)}. \tag{5.96}$$

(d) By comparing this form of the electron heat capacity with Drude's result $c_v = \frac{3}{2}nk_b$, can you explain why the heat capacity of the "quantum" electrons is so much smaller than the "classical" electrons?

(5.6) Electric current flow in semiconductor crystals

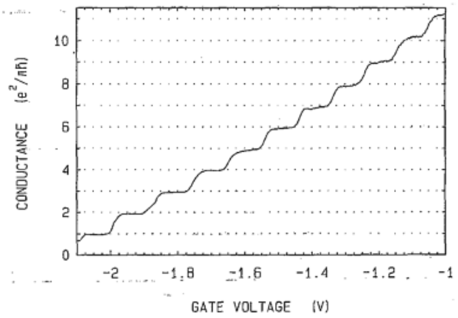


Fig. 5.24 Experimental evidence of conductance quantization measured in an AlGaAs/GaAs quantum well structure with N parallel quantized conductors from Physical Review Letters, Vol. 60, Page 848 (1998). The number of 1D conductors or subbands could be tuned by a gate voltage. We will encounter this figure again in Chapter 17, Fig. 17.4.

We have discussed that the quantum mechanical current carried by electron states is found easily by using the concept of the **group velocity** which is obtained directly from the bandstructure $E(k)$. The second important concept of current flow is how the k-states are filled/emptied by metal contacts which are reservoirs of electrons, and determine the quasi-Fermi levels of the $|k\rangle$ states in the semiconductor. In this problem, you will gain practice in applying these concepts to find the quantum mechanical electric current in various situations.

(a) Show that the ballistic conductance $G = I/V$ in any 1D metal or semiconductor at a low voltage V and at a low temperature $qV << k_bT$ is given by the *quantum of conductance* $\boxed{G_0 = g_s g_v \frac{q^2}{h}}$, where g_s is the spin degeneracy and g_v the valley degeneracy of the 1D band, q is the electron charge, and h is Planck's constant. In your derivation, show that rather remarkably this result is *independent* of the exact nature of the bandstructure! If there are N parallel 1D conductors or N uncoupled 1D bands, then the total conductance is NG_0. This quantization is indeed measured experimentally, as seen in Fig. 5.24.

(b) Electrons sit in the n_z state of a heterostructure 2D quantum well of length L_z in the z-direction and infinite depth, and are free to move in an area L_xL_y in the $x-y$ directions in the well. The energy bandstructure is $E(k_x, k_y) = \frac{\hbar^2(k_x^2+k_y^2)}{2m^\star}$. Show that the probability current density for state $|\mathbf{k}\rangle = (k_x, k_y, k_{n_z})$ is the following:

$$\mathbf{j}(k_x, k_y, k_{n_z}) = \frac{1}{L_xL_y} \cdot [\frac{\hbar}{m^\star}(k_x\hat{x} + k_y\hat{y})] \cdot \frac{2}{L_z}\sin^2(k_{n_z}z). \tag{5.97}$$

(c) Provide an expression for k_{n_z} and explain the result. Integrate the z-component to show that the 2D probability current is in the form $\mathbf{j}_{2d}(\mathbf{k}) = \frac{1}{L^d}\mathbf{v}_g(\mathbf{k})$, where $\mathbf{v}_g(\mathbf{k}) = \frac{1}{\hbar}\nabla_k E(\mathbf{k})$ is the group velocity. This is a more general result that applies also for particles that may appear "massless".

(d) Now fill the quantum well (k_x, k_y) states so that the 2D carrier sheet density is n_s. This defines a Fermi level E_F: find how n_s is related to E_F at any temperature. Also, what happens to E_F if $\hbar \to 0$? Why?

(e) Find the current per unit width flowing in the $+x$ direction as a function of temperature and n_s. Assume a spin degeneracy g_s and valley degeneracy g_v for each $\mathbf{k}$-state. This is of course equal to the current flowing in the $-x$ direction. Find the magnitude of the current density for $m^\star \sim 0.2m_0$ and $n_s \sim 10^{13}/\text{cm}^2$, $g_s = 2$, and $g_v = 1$ (these values are typical for GaN transistors; g_v is larger for silicon MOSFETs).

(f) In some 2D semiconductors such as graphene, the energy bandstructure is not parabolic, but conical: $E(k_x, k_y) = \hbar v_F\sqrt{k_x^2 + k_y^2}$, where v_F is a characteristic "Fermi" velocity. Find the $+x$ directed current per unit width for graphene as a function of temperature. Find the magnitude of the current per unit width for $v_F \sim 10^8$ cm/s, $n_s \sim 10^{13}/\text{cm}^2$, and $g_s = 2$, and $g_v = 2$, the values for 2D graphene.

(5.7) To gap or not to gap, that is the question

A fictitious 2D semiconductor has an energy dispersion $E(\mathbf{k}) = \pm\hbar v_F\sqrt{|\mathbf{k}|^2 + k_0^2}$, where $\mathbf{k} = (k_x, k_y)$, $\hbar$ is the reduced Planck's constant, v_F a characteristic velocity, and k_0 is a constant. Assume a spin degeneracy g_s and valley degeneracy g_v.

(a) Find the energy bandgap E_g, and show that the effective mass m_{xx} at the band edge is related to the bandgap by $2m_{xx}v_F^2 = E_g$. Argue that if k_0 is externally tunable with a gate, the bandstructure of this material can be made to look like graphene and its gap can be made zero.

(b) Find and sketch the density of states.

(c) Find the group velocity $\mathbf{v}_g(\mathbf{k})$ of the state $\mathbf{k}$. Sketch the magnitude and direction of the group velocities in the $\mathbf{k}$-space.

(d) Now the conduction band of this 2D semiconductor sheet is connected by ohmic contacts to a source and a drain. A voltage V is applied across these two terminals. Set up the expressions that will give the ballistic current in response to the applied voltage.

(e) The temperature is held very low, $T \to 0$ K. Estimate the ballistic current as a function of voltage if $qV >> \hbar v_F k_0$.

Actually this is not fictitious any more - recently 2D crystal semiconductors with a broken symmetry in the direction perpendicular to the 2D plane have been discovered such as silicene and germanene - atomicaly thin versions of silicon and germanium that have this sort of a bandstructure. They have rather interesting *toplogical* properties.

(5.8) **Been there, done that: ballistic current flow and current saturation**

We discussed the carrier density and the ballistic current flow for transport in $d = 1, 2, 3$ dimensions in this chapter. This problem is for you to practice the self-consistent calculations, understand current saturation, and connect the free-electron model to the case of ballistic transport in semiconductors semiconductors.

(a) Consider the 1D case, encountered in semiconductor quantum wires and carbon nanotubes. Use a conduction band effective mass of $m_c^\star = 0.2m_e$, valley degeneracy $g_v = 1$ and spin degeneracy $g_s = 2$. Calculate and plot the source quasi Fermi level at the source injection point $(E_{Fs} - E_c)$, and the drain quasi Fermi level $(E_{Fs} - qV - E_c)$ as a function of the voltage $0 < V < 2$ Volt for a high electron density $n_{1d} = 5 \times 10^6/\text{cm}$ at room temperature $T = 300$ K, and explain the plot. Next, plot the ballistic currents vs voltage for electron densities $n_{1d} = 1, 2, 3, 4, 5 \times 10^6/\text{cm}$, both at 300 K and at 77 K for $0 < V < 2$ Volt. You should observe that in 1D semiconductors the ballistic current does *not* depend on the 1D density at low voltages and low temperatures – why is this so? Why does the current saturate at high voltages?

(b) Now consider the 2D case, which is encountered in silicon transistors, III-V quantum well high-electron mobility transistors, and 2D crystal semiconductors and metals. For this problem, use a conduction band effective mass of $m_c^\star = 0.2m_e$, valley degeneracy $g_v = 1$ and spin degeneracy $g_s = 2$. Calculate and plot the source quasi Fermi level at the source injection point $(E_{Fs} - E_c)$, and the drain quasi Fermi level $(E_{Fs} - qV - E_c)$ as a function of the voltage $0 < V < 2$ Volt for a high electron density $n_{2d} = 5 \times 10^{13}/\text{cm}^2$ at room temperature $T = 300$ K, and explain the plot. Next, plot the ballistic current per unit width vs. voltage for electron densities $n_{1d} = 1, 2, 3, 4, 5 \times 10^{13}/\text{cm}^2$, both at 300 K and at 77 K for $0 < V < 2$ Volt. You should observe that unlike in 1D semiconductors, the ballistic current per unit width in 2D *does* depend on the 2D density at low voltages and low temperatures – why is this so? Why does the current saturate at high voltages?

In this problem, you have solved the 1D and 2D ballistic transistor problem in disguise. The variation of the carrier density in a transistor is done with a third terminal called the gate, which is the topic of Chapter 18: that is the only new thing you will do there, the current is already covered in this chapter.

(5.9) **Velocity saturation in nanotubes and semiconductors**

(a) If high energy electrons collide with the lattice and emit optical phonons at a very fast rate, they come to equilibrium with the lattice rather than the source and drain electrodes. Assume we have a metallic carbon nanotube that has a 1D energy dispersion $E(k) = \hbar v_F |k|$ with a spin degeneracy of $g_s = 2$ and a valley degeneracy $g_v = 2$. Show that if the optical phonon energy is $\hbar\omega_{op}$ and the above ultrafast optical phonon emission occurs, then the saturation current in the nanotube is given by $I_{sat} = \frac{q g_s g_v \omega_{op}}{2\pi}$. Find the magnitude of this current for $\hbar\omega_{op} \sim 160$ meV, and compare with experimental data (give references).

(b) At low electric fields, the velocity v of electrons in a semiconductor increases linearly with the field F according to $v = \mu F$, where $\mu = q\tau_m / m^\star$ is the mobility, τ_m is the momentum scattering time and $m^\star$ the electron effective mass. But when the electric field is cranked up, the electron velocity saturates, because the electrons emit optical phonons each of energy $\hbar\omega_{op}$ every τ_E seconds, dumping the energy qFv they gain from the electric field every second. Setting up the equations for the conservation of momentum and energy, and solving for the steady state yields an estimate of this saturation velocity. Show that the saturation velocity obtained by this scheme is $v_{sat} = \sqrt{\frac{\hbar\omega_{op}}{m^\star}} \cdot \sqrt{\frac{\tau_m}{\tau_E}}$. Show that for a typical semiconductor for which $\hbar\omega_{op} \approx 60$ meV, $m^\star \sim 0.2 m_e$, and $\tau_m \sim \tau_E$, the electron saturation velocity is of the order of $\sim 10^7$ cm/s. This is a good rough number for the saturation velocity of most semiconductors.

(5.10) Fermi energy, average energy, and conductivity of graphene

The energy-momentum dispersion relationship for electrons in 2-dimensional graphene is $E(k_x, k_y) = \hbar v_F \sqrt{k_x^2 + k_y^2}$, where $v_F = 10^6$ m/s is a parameter with dimensions of velocity. Each allowed k-state holds $g_s = 2$ spins, and there are $g_v = 2$ valleys. Using this energy dispersion, the density of states for graphene is found (see Exercise 5.2) to be $g(E) = \frac{g_s g_v}{2\pi(\hbar v_F)^2}|E|$.

(a) If the sheet of graphene has $n_{2d} = 10^{13}/\text{cm}^2$ electrons, find the Fermi wavevector k_F at $T = 0$ K.

(b) Find the Fermi energy E_F at $T = 0$ K. Express its numerical value in eV units.

(c) Find the average energy of the electron distribution at $T = 0$ K.

(d) Find an expression for the group velocity of electrons occupying the state (k_x, k_y). Remember that the group velocity is a vector in 2D.

(e) Make a sketch of the direction and magnitude of the group velocity of a few (k_x, k_y) points using part (d). How is this case different from a free-electron?

(f) When I apply an electric field, a current will flow in graphene. Sketch how the **k**-space picture changes from the equilibrium value to support the flow of current.

(5.11) Semiconductor density of states

Electrons in a very narrow bandgap 3D semiconductor are found to have a bandstructure $E(k_x, k_y, k_z) = \hbar v_F \sqrt{k_x^2 + k_y^2 + k_z^2}$, where $\hbar$ is the reduced Planck's constant, and $v_F = 10^8$ cm/s is a characteristic Fermi velocity. Assume the spin degeneracy of each **k**-state to be $g_s = 2$ and the valley degeneracy to be $g_v = 1$ for numerical evaluations, but first give the analytical answers.

(a) Show that the DOS per unit volume is $g(E) = \frac{g_s g_v}{2\pi^2} \frac{E^2}{(\hbar v_F)^3}$.

(b) Sketch the DOS and a Fermi occupation function with the Fermi level at $E = 0$, and qualitatively indicate in the sketch how the occupied electron states would change with temperature.

(c) The electron density at a temperature T is given by $n = \int dE \cdot g(E) \cdot f(E)$, where $f(E)$ is the Fermi–Dirac function characterized by a Fermi level E_F. Assume that the Fermi level is at $E_F = 0$ eV. Show that at temperature T, there are $n_{3d} = \frac{3\zeta(3)}{2} \frac{g_s g_v}{2\pi^2} (\frac{k_b T}{\hbar v_F})^3$ electrons occupied. Check the units. In other words, show that the electron density increases as T^3. Find the 3D electron density at room temperature, $T = 300$ K. [You are given that $\int_0^\infty dx \frac{x^2}{1+e^x} = \frac{3}{2}\zeta(3)$, where $\zeta(...)$ is the zeta function, and $\zeta(3) \approx 1.2$.]

(d) Indicate the group velocity vectors $\mathbf{v}_g(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(k)$ of the **k**-states and show that they have the same magnitude at all points in the **k**-space, pointing radially outwards from the Γ-point, which is $\mathbf{k} = (0, 0, 0)$. Discuss why the velocity vector is not well defined *only at the origin*.

(5.12) Quantum 2D free electrons in a magnetic field

This problem is kindly provided by Prof. Farhan Rana. Consider a 2D free electron gas confined to the x-y plane. In the Sommerfeld model, the energy of an electron with wavevector $\mathbf{k}$ is $E(\mathbf{k}) = \frac{\hbar^2 k^2}{2m_e}$, and the velocity is $\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}) = \frac{\hbar\mathbf{k}}{m_e}$. Now suppose a DC magnetic field $\mathbf{B} = B_0\mathbf{z}$ is switched on in the z-direction, as shown in Fig. 5.25 (a). In the presence of the magnetic field, because of the Lorentz force, the momentum of the electron satisfies the equation (assuming no electric field and no scattering)

$$q\mathbf{v}(\mathbf{k}) \times \mathbf{B} = \hbar \frac{d\mathbf{k}}{dt}, \tag{5.98}$$

which is the quantum version of Newton's law, with the Lorentz force.

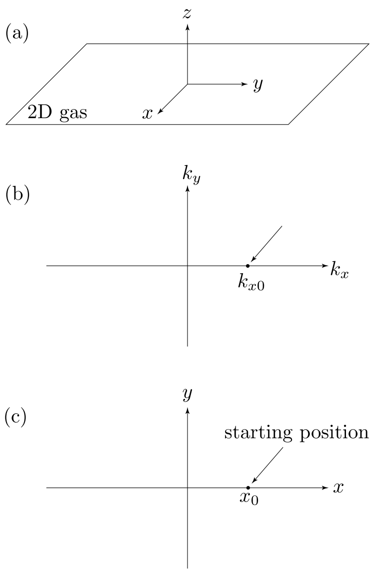


Fig. 5.25 (a) 2D electron in a perpendicular magnetic field. (b) 2D electron *k*-space. (c) 2D electron real space.

(a) In the k-space, if the starting position of the electron (before the magnetic field was switched on) is at $(k_{x0}, 0))$ as shown in Fig. 5.25 (b), then find the trajectory of the electron in the **k**-space after the magnetic field has been switched on. Plot the trajectory in the **k**-space.

(b) Continuation of part (a): If in addition, the starting position of the electron (before the magnetic field was switched on) in real space is at $(x_0, 0)$ as shown in Fig. 5.25 (c), then find the trajectory of the electron in real-space after the magnetic field has been switched on and plot it in the real space.

(c) If you did parts (a) and (b) correctly, you would have found that the motion of electron in both k-space and real space is *periodic*. Find the time period for the motion (i.e. the time taken by the electron to complete one period).

(d) Staring from the Equation 5.98, prove that the energy of the electron is conserved (i.e. does not change) during its motion. Hint: The proof is just two lines of math.

(e) If instead of one electron, there were many, before the magnetic field was switched the total current carrier by the electron gas (summing up contributions from all electrons) was given by

$$\mathbf{J} = 2q \int \frac{d^2k}{(2\pi)^2} f(\mathbf{k})\mathbf{v}(\mathbf{k}) = 0, \tag{5.99}$$

where $f(\mathbf{k})$ was the equilibrium Fermi–Dirac distribution for electrons. Find the total current carried by the electron gas *after* the magnetic field has been switched on and explain your answer.

(5.13) Particle in a box

The particle in a box problem in Section 5.6 of infinitely tall barriers is a powerful tool to estimate using back of the envelope techniques several quantum length, velocity, and energy scales. You can get a flavor for some by elementary analysis.

(a) The energy of the n$^{\text{th}}$ eigenstate is $E_n = \frac{n^2h^2}{8mL^2}$ for a particle of mass m in a quantum well of size L. Imagine a semi-classical situation in which instead of of quantum confinement, the discrete energies are because the particle in the n$^{\text{th}}$ state is moving with a velocity v_n. Show that $v_n = \frac{nh}{2mL}$.

(b) Assume an electron of mass $m_e = 9.1 \times 10^{-31}$ kg in a well of size $L_0 = 2$ Å, representative of typical atomic radii. Show that the $n = 1$ ground state velocity is 1.6×10^6 m/s, and energy is 9.4 eV, fairly representative of the scale of electron ionization energies from atoms and solids.

(c) Now consider a proton of mass $m_p \sim 2000 m_e$, 2000 times heavier the electron, quantum-confined in the ground state inside the nucleus, to a distance: $L = \frac{L_0}{20{,}000} \sim 10^{-14}$ m, which is 20,000 times smaller than interatomic distances. Show that its effective velocity is much faster than the electron in part (b) at 1.8×10^7 m/s, and energy much higher at 1.9 MeV.

(d) The examples in (b) and (c) are not that artificial: alpha particles ejected from the nucleus of unstable atoms do move at these velocities. The high energy is representative of nuclear reactions which occur at MeVs scales of part (c), compared to the much lower energy scales of chemical reactions, which are in a few eV due to the electronic energies of part (b).

Red or Blue Pill: Befriending the Matrix

6

6.1 The expansion principle 124
6.2 Matrix mechanics 125
6.3 Matrices and algebraic functions 127
6.4 Properties of matrix eigenvlaues 131
6.5 Looking ahead 131
6.6 Chapter summary section 132
Further reading 132
Exercises 133

Prior to Schrödinger's differential equation form of "wave-mechanics" for finding the allowed quantum states of electrons, Heisenberg, Born, and Jordan (Fig. 6.2) had developed the first complete form of quantum mechanics, but in the form of matrices. They had named it **matrix mechanics**. With the goal to explain the experimentally observed sharp and discrete spectral lines of the hydrogen atom, Heisenberg hit upon the crucial idea that if the dynamical variables of the electron such as its location $[x]$ and its momentum $[p]$ were *matrices* instead of numbers, then its energy would be a found from a matrix eigenvalue equation, which can yield discrete transition energies. Today we all know that matrices can have discrete eigenvalues, but this connection was not clear in the 1920s when matrices were rarely used in physics. John von Neumann, who was David Hilbert's (Fig. 6.1) student, later proved the equivalence of Heisenberg's matrix mechanics, and Schrödinger's wave mechanics.

In this chapter, we solidify our notion of why matrices present a natural and unified approach to obtain quantitative solutions to the quantum mechanical problem of electrons subjected to simple and increasingly complicated potentials. The goals of this chapter are to understand:

- How does the expansion principle connect linear algebraic equations to matrices?
- What is the connection between matrices and quantum mechanical eigenstates, eigenvalues, superpositions, and operators?
- What are some special properties of matrices and how do we evaluate them in practice?

Fig. 6.1 David Hilbert, a German mathematician in Gottingen who among many other subjects, developed the idea of Hilbert spaces, which are infinite dimensional matrices with special significance in quantum mechanics. In a delightful story which remains to be confirmed, Hilbert had promised a seminar in the USA on the solution of the famous Fermat's last theorem, to which Fermat had claimed he had a proof but the margin was too small to hold it. The packed audience was disappointed that his seminar had nothing to do with Fermat's last theorem. When asked, Hilbert replied his seminar title was just in case his cross-Atlantic plane crashed.

In Chapter 5, we became acquainted with the wave-mechanics method of Schrödinger and applied it to the free electron in various dimensions, and a few other problems. In this chapter, we befriend the matrix method of solving for the quantum mechanical states and energies of electrons. For most numerical solutions, this is the method of choice. With further judicious choice, the matrix equation can give analytical solutions, as we will see in several following chapters for the electron bandstructure in periodic potentials, the situation encountered most often for semiconductors. We first motivate matrix mechanics by discussing one of the most important and least emphasized principles of quantum mechanics.

6.1 The expansion principle

Fourier's theorem mathematically guarantees that *any* well-behaved function $f(x)$ can be expressed as a sum over a *complete* set of trigonometric functions (or complex exponentials): $f(x) = \sum_k a_k e^{ikx}$. Note that *any* complete set of eigenfunctions $[\ldots, e^{ikx}, \ldots]$ works! This set has an infinite number of elements and is called the Hilbert space. In practice, we typically use a restricted set for most problems. To find the Fourier coefficients, we use the "filtering" property of complex exponentials $a_{k_n} = \int dx f(x) e^{-ik_n x}$. If we tweak the function $f(x) \to f(x) + g(x) = h(x)$, then $h(x) = \sum_k a'_k e^{ikx}$ is still a valid expansion; the Fourier coefficients will be tweaked from $a_k \to a'_k$. But note that the perturbed function can still be expanded in terms of the *original* complete set of eigenfunctions. This idea leads to the expansion principle in quantum mechanics.

Here is the **expansion principle** of quantum mechanics: Any quantum state "vector" $|\Psi\rangle$ may be expanded as a linear superposition of the eigenvectors of *any* Hermitian operator $|\Psi\rangle = \sum_n a_n |n\rangle$. For most problems, the Hermitian operator of choice is the Hamiltonian operator $\hat{H} = \frac{\hat{\mathbf{p}}^2}{2m_0} + V(\mathbf{r})$, but it need not be. We choose the Hamiltonian operator since there exist a few problems that we encountered in Chapter 5 for which we know the set of *exact* eigenvectors $[\ldots, |n-1\rangle, |n\rangle, |n+1\rangle, \ldots]$. These sets of eigenvectors are *complete*. We also discussed in Chapter 5 that this choice of eigenstates are *stationary*. For example, we found the 1D electron on a ring problem with $\hat{H} = \frac{\hat{p}^2}{2m_0}$: gave eigenvalues $E(\mathbf{k}) = \frac{\hbar^2 k^2}{2m_0}$, and corresponding eigenvectors $|k_{\rm FE}\rangle$ projected to real space $\langle x|k_{\rm FE}\rangle = \frac{1}{\sqrt{L}} e^{ikx}$. This eigenstate basis $[\ldots, e^{ikx}, \ldots]$ is complete, where k takes all allowed values.

Fig. 6.2 Pascual Jordan, who with Heisenberg and Max Born in Göttingen developed the first consistent version of quantum mechanics in its *matrix* form, and named it matrix mechanics. Jordan also did seminal work on quantum field theory. Had he not associated himself with the Nazi party, he likely would have been recognized today as well as Heisenberg and Born.

Now consider any state of a harmonic oscillator $|\psi_{\rm HO}\rangle$. The expansion principle guarantees that we can expand *any* harmonic oscillator state in the basis of the free electron $|\psi_{\rm HO}\rangle = \sum_k a_k |k_{\rm FE}\rangle$. We can do the reverse too: expand the free electron states in terms of the harmonic oscillator. This is allowed as long as the potential term in the Hermitian operator does not blow up. For example, we can expand the particle in a box states in terms of the free electron states $|\psi_{box}\rangle = \sum a_k |k_{FE}\rangle$, *but not the other way around* because the particle in a box potential blows up outside the box. This should be obvious because the eigenfunctions of the particle in a box are all **zero** outside the box, and no matter how clever one is, it is not possible to linearly combine zeroes to produce a function that takes non-zero values outside the box.

The expansion principle is the backbone of **perturbation theory**, which underpins the actual usage of quantum mechanics in semiconductor physics. In this chapter, we set up the framework for using perturbation theory by describing the matrix representation of quantum mechanics.

6.2 Matrix mechanics

Since we can express any quantum state as an expansion in the eigenvectors $|\Psi\rangle = \sum_n a_n|n\rangle$, we can arrange the expansion coefficients as a column vector

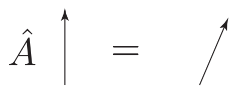
$$|\Psi\rangle = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} \langle 1|\Psi\rangle \\ \langle 2|\Psi\rangle \\ \langle 3|\Psi\rangle \\ \vdots \end{bmatrix}. \tag{6.1}$$

The **Hermitian conjugate** is obtained by transposing and taking term-by-term complex conjugation:

$$\langle\Psi| = \begin{bmatrix} a_1^\star & a_2^\star & a_3^\star & \cdots \end{bmatrix}, \tag{6.2}$$

which is a row vector. If the state $|\Psi\rangle$ is normalized, then clearly $\langle\Psi|\Psi\rangle = 1$, which requires $\sum_n |a_n|^2 = 1$. Upon measurement, the state will *always* materialize one of the eigenstates $|n\rangle$. Then $|a_n|^2$ should be interpreted as the probability the quantum state materializes in state $|n\rangle$, and a_n is the corresponding probability amplitude. The normalization condition $\sum_n |a_n|^2 = 1$ makes sure that the probabilities add up to one, and the particles are neither created nor destroyed, but their total number stays fixed. Also, the state of eigenvectors $|n\rangle$ can always be chosen to be mutually orthogonal, i. e., $\langle m|n\rangle = \delta_{mn}$. This is the basis we will work with. Also note that projecting $|\Psi\rangle = \sum_n a_n|n\rangle$ on state $\langle n|$ yields the coefficients $a_n = \langle n|\Psi\rangle$ for each n. Then, we can write the expansion as $|\Psi\rangle = \sum_n |n\rangle\langle n|\Psi\rangle$, which means that

$$\sum_n |n\rangle\langle n| = 1. \tag{6.3}$$



$$\hat{A}|\Psi\rangle = |\Phi\rangle$$

$$[A][\Psi] = [\Phi]$$

Fig. 6.3 Three ways of saying the same thing. The operator $\hat{A}$ rotates a state vector $|\Psi\rangle$ into $|\Phi\rangle$. The pictorial depiction is equivalent to the algebraic operator equation, which in turn is equivalent to the matrix form $[A][\Psi] = [\Phi]$.

This is the "closure" relation of eigenvectors that are discrete. If the eigenvectors were continuous, then the corresponding closure relation is

$$\int dx|x\rangle\langle x| = 1. \tag{6.4}$$

The fact that the two closure relations are unity allows us to insert them wherever they will help in the evaluation of matrix elements. Consider now an operator $\hat{A}$ acting on the state vector $|\Psi\rangle$. It will try to "rotate" the state vector in the Hilbert space to a state $|\Phi\rangle$ as shown pictorially in Fig. 6.3. We write this as

$$\hat{A}|\Psi\rangle = |\Phi\rangle. \tag{6.5}$$

By the expansion principle, we can expand the new state $|\Phi\rangle = \sum_m b_m|m\rangle$. Then, if we project this state on $|m\rangle$, we have

$$b_m = \langle m|\Phi\rangle = \langle m|\hat{A}|\Psi\rangle \to b_m = \sum_n a_n\langle m|\hat{A}|n\rangle = \sum_n A_{mn}a_n. \tag{6.6}$$

We see that the operator $\hat{A}$ is equivalent to a *matrix* $\hat{A} \equiv A_{mn} = [A]$. The elements of the equivalent matrix are the terms $A_{mn} = \langle m|\hat{A}|n\rangle$, obtained by the operator acting on eigenstates on both sides. We call them **matrix elements** for obvious reasons.

For example, if we choose the momentum operator acting between states $|k\rangle$ and $\langle k'|$ of the free electron, we get

$$p_{kk'} = \langle k'|\hat{p}|k\rangle = \int dx \langle k'|x\rangle\langle x|\hat{p}|k\rangle = \hbar k \delta_{k',k}. \tag{6.7}$$

Note that the "abstract" operator $\hat{p}$ has the matrix representation $\langle x|\hat{p}|x\rangle = -i\hbar\frac{\partial}{\partial x}$ in real space. The example shows that since the free-electron energy eigenstates are simultaneously momentum eigenstates, the momentum operator acting between two eigenstates extracts the value of the momentum only if the two states are identical. This is the momentum matrix element.

One of the most important operators is the Hamiltonian operator, which "extracts" the energy of the state it is acting on. If the state $|n\rangle$ happens to be an energy eigenstate, the Hamiltonian operator extracts its energy eigenvalue: $\hat{H}|n\rangle = E_n|n\rangle$. Visualize $\hat{H}|n\rangle$ as a new vector whose "direction" is the same as the eigenvector $|n\rangle$, but the length determined by the eigenvalue E_n. So the action of the Hamiltonian operator leaves the "direction" of the eigenvector $|n\rangle$ unaffected.

If the state is not an energy eigenstate but is a linear superposition $|\Psi\rangle = \sum_n a_n|n\rangle$, then the time-independent Schrödinger equation states that $\hat{H}|\Psi\rangle = E|\Psi\rangle$, which is equivalent[1] to $\hat{H}\sum_n a_n|n\rangle = E\sum_n a_n|n\rangle$. When we project this new state vector $\hat{H}|\Psi\rangle$ on the eigenvector $\langle m|$, we get an algebraic equation

$$\sum_n \langle m|\hat{H}|n\rangle a_n = E a_m, \tag{6.8}$$

for each m. Note the appearance of the matrix elements $H_{mn} = \langle m|\hat{H}|n\rangle$. If we write this out for $m = 1, 2, \ldots$, we get the set of linear equations

$$\begin{array}{rcl} H_{11}a_1 + H_{12}a_2 + H_{13}a_3 \ldots & = & Ea_1 \\ H_{21}a_1 + H_{22}a_2 + H_{23}a_3 \ldots & = & Ea_2 \\ H_{31}a_1 + H_{32}a_2 + H_{33}a_3 \ldots & = & Ea_3 \\ \vdots & = & \vdots \end{array} \tag{6.9}$$

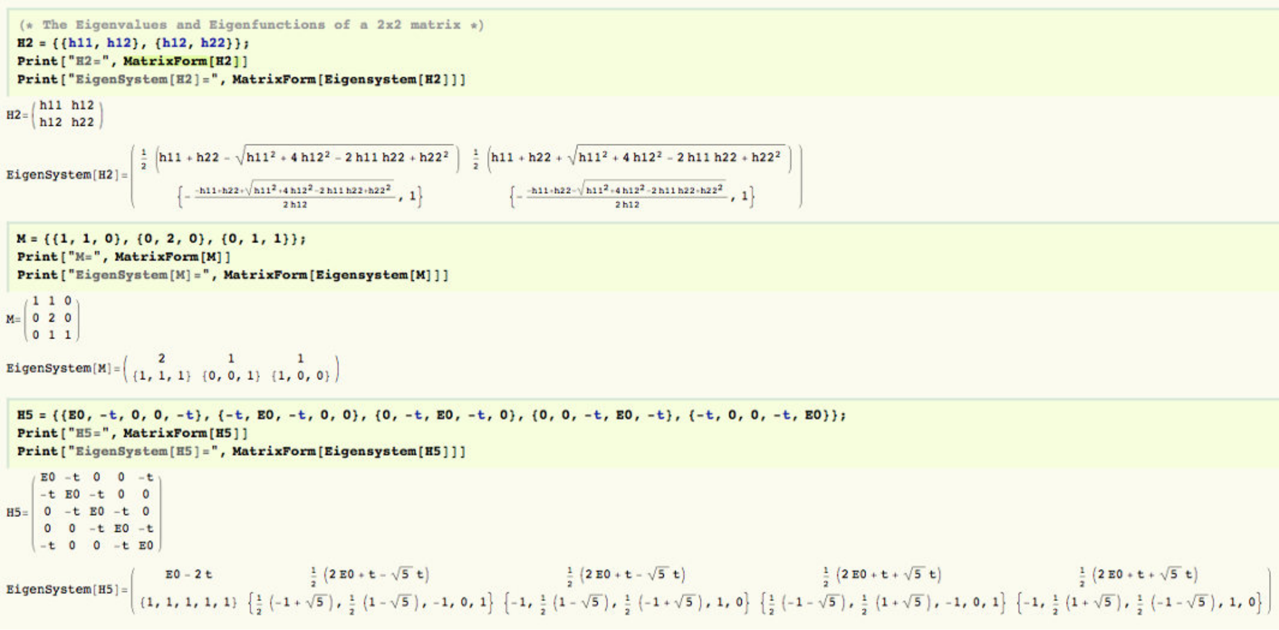
which is best captured as a matrix equation

$$\begin{bmatrix} H_{11} & H_{12} & H_{13} & \ldots \\ H_{21} & H_{22} & H_{23} & \ldots \\ H_{31} & H_{32} & H_{33} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix}. \tag{6.10}$$

Note that the Hamiltonian operator becomes a square matrix, and the state $|\Psi\rangle$ becomes a column vector. **This matrix equation contains the same information as the algebraic time-independent Schrödinger**

---

[1] This statement may be confusing: if $|\Psi\rangle$ is not an eigenstate, then $\hat{H}|\Psi\rangle = |\Phi\rangle$, $|\Phi\rangle \neq \text{constant}\cdot|\Psi\rangle$, and we cannot have $\hat{H}|\Psi\rangle = E|\Psi\rangle$, where E is a single number. This is one of the principal axioms of quantum mechanics! Note however, what is meant by the statement is that the energy E is such that it is a solution to the equation $\hat{H}|\Psi\rangle = E|\Psi\rangle$ – not all E will do, but only special ones. These special values of E in the equation $\hat{H}|\Psi\rangle = E|\Psi\rangle$ are the eigenvalues, and are what are sought for in the matrix method. The formulation is really meant for perturbation theory.



```
(* The Eigenvalues and Eigenfunctions of a 2x2 matrix *)
H2 = {{h11, h12}, {h12, h22}};
Print["H2=", MatrixForm[H2]]
Print["EigenSystem[H2]=", MatrixForm[Eigensystem[H2]]]
```

$$\text{H2=}\begin{pmatrix} h11 & h12 \\ h12 & h22 \end{pmatrix}$$

$$\text{EigenSystem[H2]=}\begin{pmatrix} \frac{1}{2}\left(h11+h22-\sqrt{h11^2+4\,h12^2-2\,h11\,h22+h22^2}\right) & \frac{1}{2}\left(h11+h22+\sqrt{h11^2+4\,h12^2-2\,h11\,h22+h22^2}\right) \\ \left\{-\frac{-h11+h22+\sqrt{h11^2+4\,h12^2-2\,h11\,h22+h22^2}}{2\,h12},\,1\right\} & \left\{-\frac{-h11+h22-\sqrt{h11^2+4\,h12^2-2\,h11\,h22+h22^2}}{2\,h12},\,1\right\} \end{pmatrix}$$

```
M = {{1, 1, 0}, {0, 2, 0}, {0, 1, 1}};
Print["M=", MatrixForm[M]]
Print["EigenSystem[M]=", MatrixForm[Eigensystem[M]]]
```

$$\text{M=}\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

$$\text{EigenSystem[M]=}\begin{pmatrix} 2 & 1 & 1 \\ \{1,1,1\} & \{0,0,1\} & \{1,0,0\} \end{pmatrix}$$

```
H5 = {{E0, -t, 0, 0, -t}, {-t, E0, -t, 0, 0}, {0, -t, E0, -t, 0}, {0, 0, -t, E0, -t}, {-t, 0, 0, -t, E0}};
Print["H5=", MatrixForm[H5]]
Print["EigenSystem[H5]=", MatrixForm[Eigensystem[H5]]]
```

$$\text{H5=}\begin{pmatrix} E0 & -t & 0 & 0 & -t \\ -t & E0 & -t & 0 & 0 \\ 0 & -t & E0 & -t & 0 \\ 0 & 0 & -t & E0 & -t \\ -t & 0 & 0 & -t & E0 \end{pmatrix}$$

$$\text{EigenSystem[H5]=}\begin{pmatrix} E0-2\,t & \frac{1}{2}\left(2\,E0+t-\sqrt{5}\,t\right) & \frac{1}{2}\left(2\,E0+t-\sqrt{5}\,t\right) & \frac{1}{2}\left(2\,E0+t+\sqrt{5}\,t\right) & \frac{1}{2}\left(2\,E0+t+\sqrt{5}\,t\right) \\ \{1,1,1,1,1\} & \left\{\frac{1}{2}\left(-1+\sqrt{5}\right),\frac{1}{2}\left(1-\sqrt{5}\right),-1,0,1\right\} & \left\{-1,\frac{1}{2}\left(1-\sqrt{5}\right),\frac{1}{2}\left(-1+\sqrt{5}\right),1,0\right\} & \left\{\frac{1}{2}\left(-1-\sqrt{5}\right),\frac{1}{2}\left(1+\sqrt{5}\right),-1,0,1\right\} & \left\{-1,\frac{1}{2}\left(1+\sqrt{5}\right),\frac{1}{2}\left(-1-\sqrt{5}\right),1,0\right\} \end{pmatrix}$$

Fig. 6.4 Examples of 2x2, 3x3, and 5x5 Matrix eigenvalue and eigenfunction calculations in Mathematica. The 2x2 Hamiltonian is general and one of the most important in all of quantum mechanics. The 3x3 matrix is a numerical example, and the 5x5 matrix of a 5-site circular ring tight-binding Hamiltonian model. Note that the eigenvectors (or eigenfunction coefficients a_n are evaluated for each eigenvalue, which is *very* nice.

equation, and is well suited for actual calculations. If we choose to work with a restricted set of say 10 states, then we have a 10×10 matrix with 10 eigenvalues and their corresponding eigenfunctions. Fig. 6.4 shows a few examples of matrix evaluations using the Mathematica package.

Historically Heisenberg developed quantum mechanics in its matrix representation and called it "matrix mechanics". Schrödinger found the algebraic version which appealed more to researchers since we are trained much better in algebra since high school than in matrices. But they are one and the same thing.

6.3 Matrices and algebraic functions

Numbers are solutions to algebraic equations. We start our education learning about integers because they quantify the fingers on our hand, and soon expand into the regime of real numbers. Soon, we realize there are algebraic equations such as $x^2 + 1 = 0$ which have solutions that are not real numbers, and realize there must be new kinds of numbers. Complex numbers contain $i = \sqrt{-1}$, which (unfortunately[2]) is called an imaginary number. We learn how to add, subtract, multiply, divide, take square roots, exponentiate, etc... with numbers.

One can visualize a matrix as an extension of the concept of a "num-

[2] *"One might think ... that imaginary numbers are just a mathematical game having nothing to do with the real world. From the viewpoint of positivist philosophy, however, one cannot determine what is real. All one can do is find which mathematical models describe the universe we live in. It turns out that a mathematical model involving imaginary time predicts not only effects we have already observed but also effects we have not been able to measure yet nevertheless believe in for other reasons. So what is real and what is imaginary? Is the distinction just in our minds?"* – Stephen Hawking

ber". For example, the algebraic equation $ax = b$ has the solution $x = b/a$, a number. A *set* of algebraic equations with multiple variables written in the form $AX = B$ has the solution $X = A^{-1}B$. Somewhere along the line, if we do not use matrices, we forget their power and beauty! We get busy using algebraic equations extensively. Turns out every algebraic equation may be written as a matrix equation. Then we can use powerful theorems of matrices to solve or analyze them. Indeed, most *numerical* approaches to solving equations have to go through the matrix route. Consider the equation of a unit circle

$$x^2 + y^2 = 1. \tag{6.11}$$

This may not look like a matrix equation at all, till we define the coordinate "vector" $X = \begin{bmatrix} x \\ y \end{bmatrix}$. Its transpose is a row vector $X^T = [x, y]$, and the matrix version of the equation of the unit circle is then

$$X^T X = 1. \tag{6.12}$$

Now consider the equation

$$x^2 + axy + y^2 = 1, \tag{6.13}$$

which can be written as

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & u \\ a-u & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = X^T A X = 1. \tag{6.14}$$

This works for any value of u. So for the unit circle, $A = I$, where I is the unit matrix. The matrix A captures the *geometric* property of the curve – whether it is a circle, or a more complex shape[3].

The strangest and most striking property of matrices is that they do not necessarily *commute*. Which is to say that in general for square matrices A and B, $AB \neq BA$. As a mathematical object, therefore they are quite distinct from real or complex numbers. Matrices thus form the natural objects for non-commutative algebra. Therefore they are central to the tenets of quantum mechanics, which is built upon the non-commutativity of operators as embodied by $\hat{x}\hat{p}_x - \hat{p}_x\hat{x} = i\hbar$, which actually was derived by Heisenberg and Born for the first time in its matrix form $[x][p] - [p][x] = i\hbar[I]$.

A square matrix A has eigenvalues λ_i and eigenvectors $[x_i]$ which are obtained by solving the equation

$$A[x] = \lambda[x] \rightarrow [A - \lambda I][x] = 0. \tag{6.15}$$

After finding the eigenvalues and eigenvectors, the square matrix can be re-written in it's **spectral decomposition**

$$A = UDU^{-1}, \tag{6.16}$$

where D is a diagonal matrix of eigenvalues

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{6.17}$$

---

[3] Such decompositions help understand the *curvature* of a function. For example, in the calculation of bandstructures of crystals, A will take the form of a Hessian matrix, and the "curvature" and topology of bands and their coresponding allowed eigenvalues and eigenstates can be concluded by examining the property of the matrices.

and the unitary transformation matrix U is formed by arranging the eigenvectors in the same order as the eigenvalues

$$U = \begin{bmatrix} [x_1] & [x_2] & \cdots \end{bmatrix}. \tag{6.18}$$

Note that U is invertible, meaning it's determinant cannot be zero. Now let's say the square matrix A is actually the Hamiltonian matrix of a quantum mechanics problem. Then solving the time-independent Schrödinger equation is *equivalent* to diagonalizing the Hamiltonian matrix by solving the matrix equation

$$\begin{bmatrix} H_{11} & H_{12} & H_{13} & \dots \\ H_{21} & H_{22} & H_{23} & \dots \\ H_{31} & H_{32} & H_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix}. \tag{6.19}$$

Clearly, it is equivalent to

$$\begin{bmatrix} H_{11} - E & H_{12} & H_{13} & \dots \\ H_{21} & H_{22} - E & H_{23} & \dots \\ H_{31} & H_{32} & H_{33} - E & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{bmatrix} = 0. \tag{6.20}$$

If instead of the infinite matrix, we choose a restricted eigenbasis set of N, then the solutions of the corresponding algebraic equation $\text{Det}[H - EI] = 0$ yield N eigenvalues E_n. Corresponding to each eigenvalue E_n, there is an eigenvector $|n\rangle$, which is a column vector. We then construct the unitary operator U by collecting the eigenvectors and write the Hamiltonian matrix in its diagonal form as

$$\begin{bmatrix} H_{11} & H_{12} & H_{13} & \dots & H_{1N} \\ H_{21} & H_{22} & H_{23} & \dots & H_{2N} \\ H_{31} & H_{32} & H_{33} & \dots & H_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H_{N1} & H_{N2} & H_{N3} & \dots & H_{NN} \end{bmatrix} = U \begin{bmatrix} E_1 & 0 & 0 & \dots & 0 \\ 0 & E_2 & 0 & \dots & 0 \\ 0 & 0 & E_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & E_N \end{bmatrix} U^{-1} \tag{6.21}$$

This is the **spectral decomposition** of the Hamiltonian $H = UDU^{-1}$, where D is a diagonal matrix whose elements are the energy eigenvalues. The exact solution requires the matrices to be infinite-dimensional, but for most practical cases we work with a restricted set.

Now let's imagine that the Hamiltonian matrix is perturbed to $H \to H_0 + W$. The eigenvalues and eigenfunctions will change. But the expansion principle tells us that the new state vector of the perturbed system can still be expanded in terms of the unperturbed eigenvectors, or the matrix U. The matrix formalism makes such perturbations easy to deal with. We will return to this problem in the next chapter.

The sum of the diagonal elements of a square matrix is called its trace, $\text{Tr}[H] = \sum_n H_{nn}$. For square matrices A, B, $\text{Tr}[AB] = \text{Tr}[BA]$.

Thus, we get $\mathrm{Tr}[H] = \mathrm{Tr}[UDU^{-1}] = \mathrm{Tr}[U^{-1}UD] = \mathrm{Tr}[D] = \sum_n E_n$. **The trace of the Hamiltonian is the sum of its eigenvalues**.

The quantum states are represented as column vectors $|\Psi\rangle = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}$, as discussed earlier. Consider another quantum state $|\Phi\rangle = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \end{bmatrix}$. If we take the projection $\langle\Phi|\Psi\rangle = \sum_n a_n b_n^\star$, we get a *number*. This is analogous to taking the dot product of two vectors, and is called the **inner product** for Dirac notation. But we can also take an **outer product**

$$|\Psi\rangle\langle\Phi| = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \begin{bmatrix} b_1^\star & b_2^\star & \dots & b_N^\star \end{bmatrix} = \begin{bmatrix} a_1 b_1^\star & a_1 b_2^\star & \dots & a_1 b_N^\star \\ a_2 b_1^\star & a_2 b_2^\star & \dots & a_2 b_N^\star \\ \vdots & \vdots & \ddots & \vdots \\ a_N b_1^\star & a_N b_2^\star & \dots & a_N b_N^\star \end{bmatrix}, \tag{6.22}$$

which is no longer a number but a *matrix*. This matrix clearly retains the information of the phases of each quantum state and their respective projections on eigenstates. This information is lost when the inner product is taken. The outer product leads to the concept of **density matrices**, which keep track of the phase relationships and interferences of quantum states[4]. An interesting result is that the trace of the inner and outer products are the same, i.e. $\mathrm{Tr}[|\Psi\rangle\langle\Phi|] = \mathrm{Tr}[\langle\Phi|\Psi\rangle] = \sum_n a_n b_n^\star$. The mathematical name for objects like the outer product $|m\rangle\langle n|$ is a *dyadic*, it is a tensor constructed out of two vectors.

That the outer product is a matrix implies we can think of it as an *operator*. In fact, we can construct operators of the form $\hat{A} = \sum a_n |n\rangle\langle n|$ with suitable coefficients a_n. One such construction will prove rather useful. Consider the Schrödinger equation $\hat{H}|n\rangle = E_n|n\rangle$ with eigenvalues E_n and eigenvectors $|n\rangle$. We define a new operator

$$\hat{G}(E) = \sum_n \frac{|n\rangle\langle n|}{E - E_n}, \tag{6.23}$$

where the coefficients are $a_n = 1/(E - E_n)$, with units of inverse energy. This operator is called the **Green's function operator**. We will use it in the later chapters. For now, note that it blows up every time $E = E_n$, and changes sign as E crosses E_n. Note what happens when the Green's function operator acts on an eigenstate $|m\rangle$:

$$\hat{G}(E)|m\rangle = \sum_n \frac{|n\rangle\langle n|}{E - E_n}|m\rangle = \sum_n \frac{|n\rangle}{E - E_n}\langle n|m\rangle = \frac{1}{E - E_m}|m\rangle. \tag{6.24}$$

This happens for every eigenvalue, because $|m\rangle$ can be any of the eigenvectors. We can picturize the Green's function operator as a "energy-eigenvalue detector". As we sweep the energies, every time $E = E_m$, there is a very large response since $\hat{G}(E_m)|m\rangle \to \pm\infty$. The

[4] Density matrices prove useful to track the evolution of quantum states in various schemes of quantum computation.

response is low when $E \neq E_m$. This is analogous to the concept of a "impulse response" in linear systems.

The Schrödinger equation may be written as $(E - \hat{H}^0)|\psi\rangle = 0$. Then note that if we apply the Green's function operator on the left side of the equation, we get

$$\hat{G}(E)(E - \hat{H}^0)|\psi\rangle = \sum_n \frac{|n\rangle\langle n|}{E - E_n}(E - \hat{H}^0)|\psi\rangle = \sum_n |n\rangle\langle n|\psi\rangle = |\psi\rangle. \tag{6.25}$$

From the above, it is clear that $\hat{G}(E) = (E - \hat{H}^0)^{-1}$, i.e., the Green's function operator is the inverse operator of $(E - \hat{H}^0)$. You can think of this in terms of matrices to make it more concrete[5].

6.4 Properties of matrix eigenvlaues

A few properties of eigenvalues of square matrices that will prove useful in later chapters are summarized here without proof. They are:

(1) If a $N \times N$ matrix M is equal to its complex conjugate transpose $M = [M^\star]^T$, then all its N eigenvalues are real. Such matrices are called Hermitian, and this is the famous rule that Hermitian matrices are guaranteed to have real eigenvalues.

(2) The determinant of a matrix is equal to the product of its eigenvalues, i.e.,

$$\boxed{\mathrm{Det}[M] = \lambda_1 \lambda_2 ... \lambda_N.} \tag{6.26}$$

(3) The order of a product of matrices inside the determinant does not matter (in fact, the matrices A and B do not have to be square matrices either!)

$$\boxed{\mathrm{Det}[AB] = \mathrm{Det}[BA].} \tag{6.27}$$

(4) The sum of the diagonal elements, or the Trace of a matrix is equal to the sum of its eigenvalues[6], i.e.,

$$\boxed{\mathrm{Trace}[M] = \lambda_1 + \lambda_2 + ... \lambda_N.} \tag{6.28}$$

(5) Even if matrices do not commute, their trace does commute, just like the determinant, and the matrices do not have to be square:

$$\boxed{\mathrm{Trace}[AB] = \mathrm{Trace}[BA].} \tag{6.29}$$

6.5 Looking ahead

What is the advantage of the spectral decomposition of a matrix $A = UDU^{-1}$? Let's observe what happens when we try to square the matrix A.

---

[5] Also, the right side of the Schrödinger equation was zero, meaning $\hat{G}(E)0 = |\psi\rangle$. This may seem weird because the Green's function operator seems to act on "zero" to create the state $|\psi\rangle$. We will return to this strange behavior in Chapter 13 to explore what it is trying to say.

[6] When we evaluate the electron bandstructure of semiconductors in later chapters, this mathematical property translates to physical **sum rules** for eigenvalues in bands.

$$A^2 = UD(U^{-1}U)DU^{-1} = UD^2U^{-1} = U \begin{bmatrix} \lambda_1^2 & 0 & \cdots \\ 0 & \lambda_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} U^{-1}. \quad (6.30)$$

$U^{-1}U = I$ contracts the expansion, and we only have to square the diagonal matrix, which is trivial since we just square the eigenvalues! Think of any higher order power of the matrix: the U's always contract, so $A^N = UD^NU^{-1}$! If we visualize matrices as extensions of real and complex numbers, we should be curious about doing similar operations on them. For example, what is the square root of a matrix? What is the exponential or logarithm of a matrix? Can we take sines and cosines of matrices? The answer to all of these questions is yes, and the spectral decomposition is the first step to such fun. In this chapter, we discussed the *linear* properties of matrices, which will help us get started with perturbation theory.

6.6 Chapter summary section

In this chapter, we learned:

- Quantum states can be represented by analytical wavefunctions of the form $\psi(x) = \sum_n a_n \psi_n(x)$, or equivalently as the column vectors $[\psi] = [..., a_n, ...]^T$. The Schrödinger equation in the differential form $[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)]\psi(x) = E\psi(x)$ is equivalent to Heisenberg's matrix form $[H][\psi] = E[\psi]$.
- Physical observables that are represented by operators in Schrödinger's formulation are represented by corresponding matrices in Heisenberg's matrix formulation. One can visualize the matrix of the physical observable $[A]$ as an operator acting on a quantum state vector $[\psi]$, and giving a new state vector $[\psi']$, i.e. $[A][\psi] = [\psi']$. The eigenvalues and eigenfunctions of the matrix are the eigenvalues and eigenfunctions of the quantum system for that physical observable, i.e. $[A][\psi_A] = A[\psi_A]$.
- Several properties of matrices and their computer implementation will be used for the solution of quantum mechanical problems that present difficulties in the analytical formulation.

Further reading

Most books on electronic properties of solids or quantum mechanics stick to the differential equation approach and do not embrace matrices. Unless you are a math major or a math aficionado, you typically see matrices in introductory math courses, and then grow comfortable with them by applying them in courses in science and engineering. There are too many good books on matrices and linear algebra, so I will be brief. To learn the properties of matrices to your heart's content, I recommend **Linear Algebra and its Applications** by G. Strang. **Introduction to Fourier Analysis and Generalised Functions** by M. Lighthill presents a solid foundation on Fourier series that connects to Dirac's insights to its connection to quantum mechanics. Max Born and Heisenberg had developed the first consistent form of quantum mechanics in the matrix form; the book **Quantum Mechanics in Simple Matrix Form** by T. Jordan explains quantum mechanics from the matrix perspective in an approachable and illuminating way.

Exercises

(6.1) **Some properties of matrices**

(a) The trace of a square matrix is the sum of its diagonal elements. Show that though two square matrices A and B may not commute $AB \neq BA$, their trace does: $\text{Trace}(AB) = \text{Trace}(BA)$.

(b) Using the spectral decomposition of a square matrix $A = UDU^{-1}$ and part (a) of this problem, show that the trace of a square matrix is equal to the sum of all the eigenvalues E_n of the square matrix, $\text{Trace}(A) = \sum_n E_n$.

(c) For any complex number z, the Taylor series of its exponent is $e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + ...$ Using the spectral decomposition of a square matrix $A = UDU^{-1}$, show that a square matrix A follows the same exponential relation as a complex mumber: $e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + ...$

(d) For two complex numbers z_1 and z_2, $e^{z_1}e^{z_2} = e^{z_1+z_2}$. Verify that the matrix version of this relation $e^A e^B = e^{A+B+[A,B]/2+...}$, where $[A, B] = AB - BA$ is the commutator of the matrices A and B. Several commutators of the matrices A and B show up in the exponent. This is the Baker–Campbell–Hausdorff formula that plays an important role in distinguishing quantum phenomena from classical phenomena for which the commutators are always zero.

(6.2) **Using matrices to solve energy eigenvalue problems**

The next few chapters show that the quantum mechanical problem of an electron in a semiconductor crystal maps exactly to the following matrix problem. In this problem, we get comfortable in solving energy eigenvalue problems on the computer using packages (for example, see Fig. 6.4). Include plots in your solution whenever applicable.

(a) As a warm-up, re-do the examples in Fig. 6.4 in your program of choice. For example, if you are using Mathematica, get comfortable with the Eigenvalue and Eigenfunction commands. For other programs, identify the corresponding commands and obtain the solutions of Fig. 6.4.

(b) Consider N identical atoms, each of which has an allowed electron energy eigenvlaue E_0 when the atoms are far apart and do not interact. But when the atoms are arranged in a 1D periodic crystal (say labelled $1, 2, .., n, ..., N$), an electron can hop from an atom to its neighboring atoms. Let's assume that the electron only hops from an atom to its nearest neighbors to the left and the right, i.e. from $n \to n \pm 1$. In doing so, it can lower its energy by t_0. Show why the Hamiltonian of this crystal is then the $N \times N$ square matrix

$$\begin{bmatrix} E_0 & t_0 & 0 & \dots & 0 \\ t_0 & E_0 & t_0 & \dots & 0 \\ 0 & t_0 & E_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & E_0 \end{bmatrix}. \tag{6.31}$$

Note that the original (or unperturbed) eigenvalues E_0 sit in the diagonal terms, and the hopping terms t_0 sit in bands next to the diagonal terms.

Argue from physical grounds why the sign of the hopping terms should typically be negative.

(c) Implement the tightbinding matrix of Equation 6.31 on the computer for $N = 2$. Call it "tightbinding2by2". Find its eigenvalues. How many eigenvalues are there? Verify this by calculating the eigenvalues by hand.

(c) Implement the tightbinding matrix of Equation 6.31 on the computer for $N = 3$. Call it "tightbinding3by3". Find its eigenvalues. How many eigenvalues are there?

(d) Scale the implementation to $N = 100$, and call the matrix "tightbinding100by100". Use $E_0 = 1$ eV and a hopping term t_0 that you can manipulate. What are the N eigenvalues when $t_0 = 0$? Is it reasonable? This collection of eigenvalues is called the **bandstructure**, just that for $t_0 = 0$ there is not much structure to it!

(e) Using your matrix "tightbinding100by100", plot the bandstructure for a few t_0 by manipulating its value. What is the width of the band (i.e. the difference of the maximum and minimum eigenvalues)? For $t_0 = -0.3$ eV, if I put $N/5 = 20$ electrons into the crystal, where will the Fermi energy E_F be at $T \to 0$ K?

(f) Modify "tightbinding100by100" to allow for electrons to hop to the next-nearest neighbors, meaning from atom number $n \to n \pm 2$, with a slightly lower strength t_1. How does the bandstructure change?

(6.3) **The Pauli spin matrices**
The four 2×2 matrices

$$\sigma_0 = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{6.32}$$

$$\sigma_1 = \sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{6.33}$$

$$\sigma_2 = \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \tag{6.34}$$

and

$$\sigma_3 = \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{6.35}$$

are rather special in quantum mechanics. The first is the 2×2 identity matrix, and σ_x, σ_y, and σ_z are called the Pauli spin matrices.

(a) Verify that $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = I_2$, i.e. in matrix language, the three Pauli matrices are the square roots of unity.

(b) Verify that each one of the three Pauli spin matrices are Hermitian.

(c) Now consider the Pauli matrix operator acting on a quantum state $[\psi]$ to give $\sigma_x[\psi] = [\phi]$. Acting on the same state again gives $\sigma_x\sigma_x[\psi] = \sigma_x^2[\psi] = [\psi]$. Argue why this implies the eigenvalue of σ_x^2 is $+1$, and since σ_x is Hermitian, the eigenvalues of σ_x are ± 1, and this holds for each of the three Pauli matrices.

(d) Verify that the Pauli matrices form a cyclic commutation group, i.e., $[\sigma_x, \sigma_y] = \sigma_x\sigma_y - \sigma_y\sigma_x = 2i\sigma_z$. The physical meaning of this relation: a rotation operation of a quantum state vector about the x axis followed by one about the y axis does not lead to the same state vector if the operations are done in reverse, they are off by a rotation around the z axis.

(6.4) **Fundamentals of 2 × 2 matrices**
Consider the general 2×2 Hermitian matrix

$$H_2 = \begin{bmatrix} a & b \\ b^\star & c \end{bmatrix}. \tag{6.36}$$

(a) Show that the matrix can be written as $H_2 = h_0 I_2 + \mathbf{h} \cdot \vec{\sigma}$, where $\mathbf{h} = (h_x, h_y, h_z)$ is a vector whose components are three complex numbers, and the three components of $\vec{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are the three Pauli matrices. Express h_0, h_x, h_y, h_z in terms of the matrix elements a, b, c.

(b) Because the eigenvalues of the Pauli matrices are ± 1, show that the eigenvalues of the general 2×2 matrix are $h_0 \pm |\mathbf{h}|$.

(c) The general structure of the 2×2 matrix will be of utmost importance in later chapters to explain the formation of semiconductor bandstructure, and in general quantum mechanical problems of transitions and time evolution. Explore a few properties of the eigenfunctions and eigenvalues of the matrix by looking at special cases of the matrix elements, such as $a = c$, $b = 0$, small values of b, etc.

Perturbations to the Electron's Freedom

7

7.1 Degenerate perturbation theory 136
7.2 Non-degenerate perturbation theory 138
7.3 The Brillouin–Wigner perturbation results 141
7.4 Rayleigh–Schrödinger perturbation results 142
7.5 The Hellmann–Feynman theorem 143
7.6 Perturbation theory example 144
7.7 Chapter summary section 147
Further reading 148
Exercises 149

In Chapter 5 we discussed a few *exactly* solved problems in quantum mechanics. We also found that the many applied problems may not be exactly solvable in an analytical form. The machinery to solve such problems is perturbation theory. To understand the physics of semiconductors and their nanostructures, we need the tool of perturbation theory in our arsenal. In Chapter 6, we developed the matrix formalism of quantum mechanics, which is well-suited to handle perturbation theory. Sometimes we will be able to reduce the matrix solutions to closed-form algebraic forms, which always helps in visualization. In this chapter, we develop an additional analytical tool for perturbation theory that is indispensable in the insight it provides. The goals of this chapter are then the following:

- Develop a matrix approach to perturbation theory, which will use the known eigenvalues and eigenfunctions of a quantum mechanics problem represented by the Hamiltonian $\hat{H}_0$ to identify the new eigenvalues and eigenfunctions upon a perturbation that changes the Hamiltonian to $\hat{H}_0 + \hat{W}$.
- Use some simplifications of non-degenerate energy eigenvalue systems to develop non-matrix based, algebraic results of perturbation theory.
- Become comfortable with the analytical Rayleigh–Schrödinger and Brillouin–Wigner non-degenerate perturbation theory results, and learn when to apply them, and when they are not applicable.
- Appreciate that the matrix version of perturbation theory is always applicable, both for degenerate and non-degenerate problems by applying it to examples in preparation of its use in semiconductors.

Let $\hat{H}^0$ be the Hamiltonian for the solved problem. Then the time-dependent Schrödinger equation is $i\hbar\frac{\partial}{\partial t}|\Psi\rangle = \hat{H}^0|\Psi\rangle$. The eigenstates of definite energy are also *stationary* states $\langle r|n\rangle = \psi_E(r)e^{-iE_n t/\hbar}$, where $|n\rangle$ are the eigenvectors and E_n the corresponding eigenvalues. Note that *all* the solved problems we discussed in Chapter 5 such as the harmonic oscillator or the particle in a box had *time-independent* potentials.

Many real-world situations involve time-dependent potentials. For example, imagine a field-effect transistor whose gate voltage is being modulated by an ac signal. That will create a potential variation for electrons of the form $V(r)e^{i\omega t}$. A similar variation will be experienced by electrons interacting with photons of an electromagnetic wave, or with phonons of lattice vibrations. Consider the limit of very small frequencies $\omega \to 0$, or a "DC" potential. Then, the potential only has a spatial variation. A dc voltage is not truly time-independent because it has to be turned on or off at some time. But most of the physics we are interested in this and a few following chapters happens when the perturbation has been "on" for a long time in the past, and things have reached steady-state. It is in this sense that we discuss time-independent perturbation theory. We defer explicitly time-varying or oscillatory perturbations to Chapter 20 and beyond.

7.1 Degenerate perturbation theory

The time-independent Schrödinger equation for the *solved* problem is

$$\hat{H}^0|n\rangle = E_n^0|n\rangle, \tag{7.1}$$

where $\hat{H}^0$ is the unperturbed Hamiltonian. That means we know all the eigenfunctions $|n\rangle$ and their corresponding eigenvalues E_n^0. This is shown in Fig. 7.1. Lets add a perturbation W to the initial Hamiltonian such that the new Hamiltonian becomes $\hat{H} = \hat{H}^0 + W$. The new Schrödinger equation is

$$(\hat{H}^0 + W)|\psi\rangle = E|\psi\rangle. \tag{7.2}$$

The perturbation W has changed the eigenvectors from $|n\rangle \to |\psi\rangle$. The old eigenvalues may not be eigenvalues of the new Hamiltonian. Some eigenvalues increase in energy, some decrease, and others may not be affected. This is illustrated in Fig 7.1. So we have to solve for the new eigenvalues E and obtain the corresponding eigenvectors.

At this stage, we invoke the **expansion principle** introduced in Chapter 6. It states that the *perturbed* state vector $|\psi\rangle$ can always be expressed as a linear superposition of the *unperturbed* eigenvectors $|n\rangle$, since the unperturbed eigenstates form a complete basis set. It is the same philosophy of expanding any function in terms of its Fourier components. Thus we write

$$|\psi\rangle = \sum_n a_n|n\rangle, \tag{7.3}$$

where a_n's are (in general complex) expansion coefficients. The coefficients are obtained by taking the projection $\langle m|\psi\rangle$, which yields $a_n = \langle n|\psi\rangle$. Then Equation 7.2 reads

$$\sum_n a_n(\hat{H}^0 + W)|n\rangle = E\sum_n a_n|n\rangle. \tag{7.4}$$

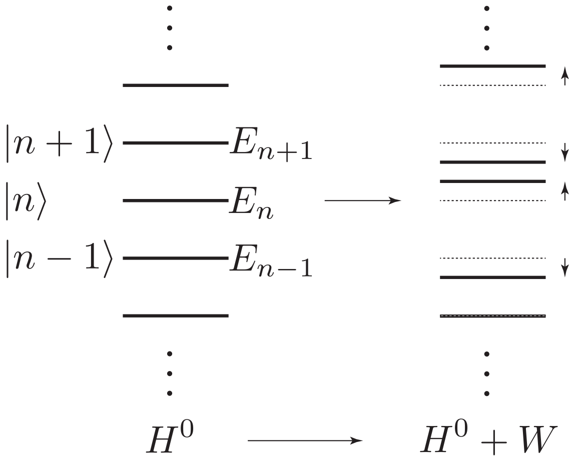


Fig. 7.1 Illustrating the effect of a perturbation. The initial eigenstates and eigenvalues of a quantum system change upon application of a perturbation W.

We can visualize the new state vector $|\Psi\rangle$ as the original eigenvector "rotated" by the perturbation W, as we had introduced in Chapter 6, and specifically in Fig. 6.3. Let us project the left and right hand sides of Equation 7.4 on $\langle m|$ to get

$$\sum_n a_n \langle m|(\hat{H}^0 + W)|n\rangle = Ea_m, \tag{7.5}$$

which is a matrix when m takes values $1, 2, \ldots N$:

$$\begin{bmatrix} E_1 + W_{11} & W_{12} & W_{13} & \ldots & W_{1N} \\ W_{21} & E_2 + W_{22} & W_{23} & \ldots & W_{2N} \\ W_{31} & W_{32} & E_3 + W_{33} & \ldots & W_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{N1} & W_{N2} & W_{N3} & \ldots & E_N + W_{NN} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix}. \tag{7.6}$$

The eigenvalues and the corresponding eigenvectors of this matrix equation are obtained by diagonalization, as discussed in Chapter 6. The new eigenvalues $E_n^{'}$ thus depend on the matrix elements $W_{mn} = \langle m|W|n\rangle$ of the perturbation. Note that if some eigenvalues of the unperturbed Hamiltonian happened to be degenerate, the matrix diagonalization method takes that into account naturally without problems. In that sense, the matrix formulation of perturbation theory is sometimes referred to as *degenerate* perturbation theory. But the matrix formulation handles non-degenerate situations equally well, and is perfectly general.

In case we did not start with a diagonal basis of the unperturbed

Hamiltonian H^0, then we have the Schrödinger equation

$$\begin{bmatrix} H^0_{11}+W_{11} & H^0_{12}+W_{12} & \dots & H^0_{1N}+W_{1N} \\ H^0_{21}+W_{21} & H^0_{22}+W_{22} & \dots & H^0_{2N}+W_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H^0_{N1}+W_{N1} & H^0_{N2}+W_{N2} & \dots & H^0_{NN}+W_{NN} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = E \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}. \tag{7.7}$$

The solutions thus reduce to diagonalizing the corresponding perturbed matrices. Many of the perturbation matrix elements $W_{mn} = \langle m|W|n\rangle$ can be made zero by suitable choice of bases, which reduces the work involved in diagonalization. Note that we can easily obtain the eigenvalues, the eigenvectors typically require more work. But with the availability of math packages such as Mathematica and MATLAB, this is done in a jiffy for most situations we will deal with.

An important question in applying degenerate perturbation theory is: which states should be included in the $N \times N$ matrix? We state the guiding principles, which will be proven in the next section on non-degenerate perturbation theory.

First principle: Eigenstates with eigenvalues widely separated in energy $E_n - E_m$ interact weakly by the perturbation.

Second principle: The change in the eigenstates energy is proportional to the perturbation matrix element squared $|W_{mn}|^2$.

Quantitatively, the perturbation in energy of a state with energy E is $\Delta E \approx |W_{mn}|^2/(E - E_n)$ by interacting with state $|n\rangle$. Thus, states for which W_{mn} terms are small or zero may be left out. If we are interested in a set of energy eigenvalues (say near the conduction and valence band edges of a semiconductor), energies far away from the band edges may also be left out. We will see the application of these rules in many following chapters.

We will see examples of degenerate perturbation theory in the next chapter (Chapter 8), where we will apply it to the problem of a free electron. That will require us to solve either 2×2 matrices, or of higher order, depending on the accuracy we need. Later on, we will also encounter it when we discuss the $\mathbf{k}\cdot\mathbf{p}$ theory of bandstructure to deal with the degeneracies of heavy and light valence bands. For now, we look at particular situations when we have "isolated" eigenvalues that are non-degenerate.

7.2 Non-degenerate perturbation theory

Schrödinger's crowning achievement was to obtain an algebraic equation, which when solved, yields the quantum states allowed for electrons. Schrödinger's equation is called the "wave"-equation because it was constructed in analogy to Maxwell's equations for electromagnetic waves. Heisenberg was the first to achieve the breakthrough in quantum mechanics before Schrödinger, except his version involved matrices which he called matrix mechanics. Because matrices were

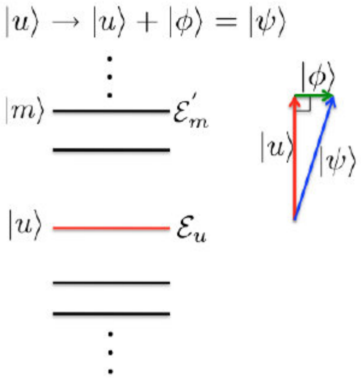


Fig. 7.2 Perturbation of vector state. The perturbation rotates the eigenvector $|u\rangle$ to $|\psi\rangle$. If we forego normalization of $|\psi\rangle$, we can find a vector $|\phi\rangle$ orthogonal to $|u\rangle$ such that $\langle u|\phi\rangle = 0$, and consequently $\langle u|\psi\rangle = 1$.

unfamiliar at that time in physics, this formulation was not immediately accepted. It was later that the mathematician von Neumann (Fig. 7.3) proved that both approaches were actually identical from a mathematical point of view.

Fig. 7.3 John von Neumann was one of the foremost mathematicians of the 20th century. He rigorously established the mathematical principles of quantum mechanics. He was instrumental in the conception and building of the first digital computer. The computer architecture of most microprocessors uses the von Neumann architecture.

So at this point, we will try to return to a "familiar" territory in perturbation theory from the matrix version presented in the previous section. We try to formulate an *algebraic* method to find the perturbed eigenvalues and eigenvectors.

Consider a perturbation $\hat{W}$ added to the solved (or unperturbed) Hamiltonian $\hat{H}^0$. The Schrödinger equation is

$$(\hat{H}^0 + \hat{W})|\psi\rangle = E|\psi\rangle, \tag{7.8}$$

and the unperturbed state $|u\rangle$ satisfies

$$\hat{H}^0|u\rangle = E_u|u\rangle. \tag{7.9}$$

The new quantum state differs from the unperturbed state, so we write

$$|\psi\rangle = |u\rangle + |\phi\rangle. \tag{7.10}$$

We can picturize the final state $|\psi\rangle$ as a "vector" sum of the unperturbed state $|u\rangle$ and the vector $|\phi\rangle$. This is schematically shown in Fig. 7.2. In particular, if we are willing to not normalize the final state, then we can always choose $|\phi\rangle$ to be orthogonal to $|u\rangle$, leading to $\langle u|\phi\rangle = 0$ and $\langle u|\psi\rangle = 1$. We can then project Equation 7.8 on $\langle u|$ to obtain the energy equation

$$E = E_u + \langle u|W|\psi\rangle = \underbrace{E_u}_{\text{unperturbed}} + \underbrace{\langle u|W|u\rangle}_{\Delta E^{(1)}} + \underbrace{\langle u|W|\phi\rangle}_{\text{higher orders}}. \tag{7.11}$$

Note that we obtain the "first-order" energy correction $\Delta E^{(1)}$: they are the diagonal matrix elements of the perturbation with the unperturbed eigenstates. Think of a "DC" perturbation – say a voltage V_0 that depends neither on space nor time. All the initial energy eigenvalues shift by the corresponding energy: $E_u \rightarrow E_u + qV_0$ due to the first order term, since $\langle u|qV_0|u\rangle = qV_0$. We will shortly see that for this particular perturbation, the higher order terms are zero because they depend on the *cross*-matrix terms of the kind $\langle m|qV_0|n\rangle = qV_0\langle m|n\rangle = 0$ for $n \neq m$. An example of such a situation is when a voltage is applied across a gate capacitor to a semiconductor – the entire bandstructure which consists of the allowed E_n's shift rigidly up or down. We call this energy band-bending in device physics. Such a "DC" perturbation does not couple different energy states for the above reason, and results in only a first-order rigid shift in energies.

Most perturbations are not of the "DC"-kind; thus we need the higher order terms for them. To do that, it is useful to define

$$E'_u = E_u + \langle u|W|u\rangle. \tag{7.12}$$

We then split the diagonal and off-diagonal elements of the perturbation just like decomposing a signal into its "DC" + "AC" terms. Think of $\hat{W}$ as an operator. Hence we are splitting a matrix into two:

$$\hat{W} = \hat{D} + \hat{W}', \tag{7.13}$$

and the total Hamiltonian then is

$$\hat{H} = \underbrace{\hat{H}^0 + \hat{D}}_{\hat{H}^{(d)}} + \hat{W}'. \tag{7.14}$$

The reason for doing this is that the unperturbed eigenvalues are going to shift by the diagonal part of the perturbation $\hat{D}$ without interacting with other states. The off-diagonal terms in $\hat{W}$ will further tweak them by interactions with other states. To move further, we write

$$(\hat{H}^{(d)} + \hat{W}')|\psi\rangle = E|\psi\rangle, \tag{7.15}$$

and rearrange it to

$$(E - \hat{H}^{(d)})|\psi\rangle = \hat{W}'|\psi\rangle, \tag{7.16}$$

At this stage, our goal is to find the perturbation vector $|\phi\rangle = |\psi\rangle - |u\rangle$. How can we obtain it from the left side of Equation 7.16 in terms of the perturbation on the right? Recall in Chapter 6 we discussed the Green's function operator briefly. We noticed that it is an "inverse" operator, meaning we expect

$$\hat{G}(E)(E - \hat{H}^{(d)})|\psi\rangle = \sum_m \frac{|m\rangle\langle m|}{E - E'_m}(E - \hat{H}^{(d)})|\psi\rangle = \sum_m |m\rangle\langle m|\psi\rangle = |\psi\rangle. \tag{7.17}$$

So to get $|\phi\rangle = |\psi\rangle - |u\rangle$, perhaps we should use the operator

$$\hat{G}(E) - \frac{|u\rangle\langle u|}{E - E'_u} = \sum_{m \neq u} \frac{|m\rangle\langle m|}{E - E'_m}. \tag{7.18}$$

Operating on the LHS of Equation 7.16 we obtain

$$\sum_{m \neq u} \frac{|m\rangle\langle m|}{E - E'_m}(E - \hat{H}^{(d)})|\psi\rangle = \sum_{m \neq u} |m\rangle\langle m|\psi\rangle \tag{7.19}$$

$$= (\sum_m |m\rangle\langle m|\psi\rangle) - |u\rangle\langle u|\psi\rangle = |\psi\rangle - |u\rangle = |\phi\rangle,$$

which is what we wanted. Now we use the same operator on the right of Equation 7.16 to finish the job. Since $\hat{W}'$ consists of only off-diagonal cross matrix elements, we write it in its outer product form as $\hat{W}' = \sum_m \sum_{n\neq m} |m\rangle\langle m|\hat{W}|n\rangle\langle n|$, and apply the "reduced" Green's function to get

$$|\phi\rangle = \sum_{l\neq u}\sum_{m}\sum_{n\neq m} \frac{|l\rangle\langle l|}{E-E'_l}|m\rangle\langle m|\hat{W}|n\rangle\langle n|\psi\rangle = \sum_{m\neq u}\sum_{n\neq m}|m\rangle\frac{\langle m|\hat{W}|n\rangle}{E-E'_m}\langle n|\psi\rangle, \tag{7.20}$$

Thus, we obtain the perturbed state $|\psi\rangle = |u\rangle + |\phi\rangle$ to be

$$|\psi\rangle = |u\rangle + \underbrace{\sum_{m\neq u}\sum_{n\neq m}|m\rangle\frac{\langle m|\hat{W}|n\rangle}{E-E'_m}\langle n|\psi\rangle}_{|\phi\rangle}. \tag{7.21}$$

As a sanity check, we note that if $\hat{W}=0$, $|\psi\rangle = |u\rangle$, as it should be. Next, we note that this is a *recursive* relation, meaning the unknown $|\psi\rangle$ also appears inside the sum on the right side. Thus, by successive substitution of the left hand side into right, it can be taken to many orders. But we are going to retain just up to the second order. That means, we will assume that the perturbation is weak, and so we are justified in replacing the $|\psi\rangle$ inside the sum on the right side by the unperturbed state $|u\rangle$.

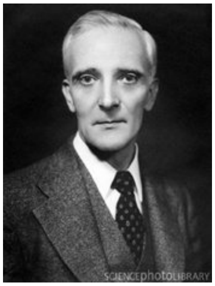


Fig. 7.4 Leon Brillouin, as in the Brillouin function, and the Brillouin zone in solid–state physics, is also the "B" of the WKB approximation. He discovered one of the three fundamental light-matter scattering processes (the other two being Rayleigh and Raman scattering).

7.3 The Brillouin–Wigner perturbation results

From the above considerations, we get the result for the perturbed eigenstates

$$\boxed{|\psi\rangle \approx |u\rangle + \underbrace{\sum_{m\neq u}\frac{\langle m|\hat{W}|u\rangle}{E-E'_m}|m\rangle}_{|\phi\rangle^{(1)}}.} \tag{7.22}$$

The perturbed state vector given by Equation 7.22 is now in a form that can be used for calculations. That is because every term on the right side is known, except for the energy E in the denominator. To obtain the perturbed eigenvalues E, we substitute Equation 7.22 into the expression for energy $E = E_u + \langle u|W|u\rangle + \langle u|W|\phi\rangle$ to obtain

$$\boxed{E \approx E_u + \underbrace{\langle u|W|u\rangle}_{\Delta E^{(1)}} + \underbrace{\sum_{m\neq u}\frac{|\langle m|\hat{W}|u\rangle|^2}{E-E'_m}}_{\Delta E^{(2)}}.} \tag{7.23}$$

This result is called the **Brillouin–Wigner** (BW) perturbation theory after those who developed it (Figs. 7.4 and 7.5). Note that we

Fig. 7.5 Eugene Wigner, a mathematical physicist who introduced seminal notions of symmetry in atomic spectra, quantum mechanics, and solid-state physics. Wigner was awarded the 1963 Nobel Prize in physics. Wigner's memorable statement: "It is nice that the computer understands the problem. But I would like to understand it too". Dirac was married to Wigner's sister.

need to solve for the unknown eigenvalues E. For multiple states, the solution would require finding the roots of a high order polynomial, since Equation 7.23 is indeed a polynomial. For example, let's say we were looking at a three-state problem with unperturbed energies E_{u1}, E_{u2}, E_{u3}, and we want to find how the eigenvalue of state $u = 2$ is modified by the perturbation. Then, the second energy energy correction has 2 terms, since $m \neq 2$. The equation then becomes a third-order polynomial with three roots, which are the desired eigenvalues.

Fig. 7.6 Lord Rayleigh was the co-discoverer of argon, and of Rayleigh scattering of long wavelength light waves from matter that explains why the sky is blue. The perturbation problem is essentially also a scattering problem in disguise, because one can imagine the unperturbed states being scattered into new states because of the perturbing potential. Rayleigh was awarded the 1904 Nobel Prize in Physics. J. J. Thomson, the discoverer of the electron, was Rayleigh's student.

7.4 Rayleigh–Schrödinger perturbation results

The BW perturbation theory requires us to solve the polynomial equations for the perturbed energies. This can be avoided if we are willing to compromise on the accuracy. If so, the unknown energy term E in the denominator of the second order correction term may be replaced by the unperturbed value, $E \to E_u$. Then the energy eigenvalues are obtained directly from

$$\boxed{E \approx E_u + \langle u|\hat{W}|u\rangle + \sum_{m \neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E_u - E'_m},} \tag{7.24}$$

and the eigenfunctions are

$$\boxed{|\psi\rangle \approx |u\rangle + \underbrace{\sum_{m \neq u} \frac{\langle m|\hat{W}|u\rangle}{E_u - E'_m}|m\rangle}_{\phi^{(1)}}.} \tag{7.25}$$

This set of perturbed eigenfunction and eigenvalues is called the **Rayleigh–Schrödinger** (RS) perturbation theory. Note that in this form, we know all the terms on the right side. It was first derived by Schrödinger right after his seminal work on the wave equation of electrons. The RS-theory is *not* applicable for understanding perturbation of *degenerate states*, as the denominator $E_n - E'_m$ can go to zero. But BW-theory applies for degenerate states too, and one can always resort back to the degenerate perturbation theory. Schrödinger originally derived this result and referred to Rayleigh's (Fig. 7.6) work on classical perturbation theory of the effect of inhomogeneities on the vibration frequencies of mechanical strings. The quantum naming scheme pays homage to the quantum and the classical versions.

In the treatment of degenerate perturbation theory earlier, we discussed the strategy to choose which states to include in the matrix. The last term in the BW or RS perturbation theory results provides the guiding principle. Note that this term goes as the perturbation matrix element *squared*, divided by the energy difference. In the absence of the perturbation, the eigenvectors corresponding to the eigenvalues were

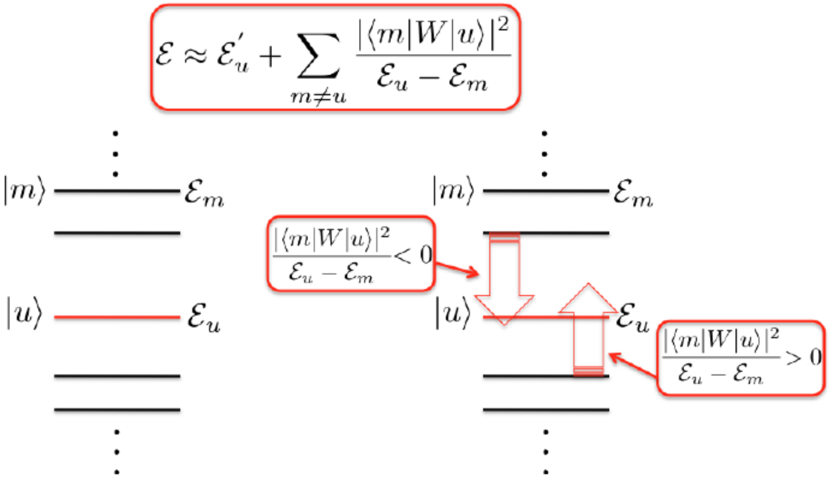

$\mathcal{E} \approx \mathcal{E}_u' + \sum_{m\neq u} \frac{|\langle m|W|u\rangle|^2}{\mathcal{E}_u - \mathcal{E}_m}$
$|m\rangle$
$\mathcal{E}_m$
$|u\rangle$
$\mathcal{E}_u$
$|m\rangle$
$\mathcal{E}_m$
$\frac{|\langle m|W|u\rangle|^2}{\mathcal{E}_u - \mathcal{E}_m} < 0$
$|u\rangle$
$\mathcal{E}_u$
$\frac{|\langle m|W|u\rangle|^2}{\mathcal{E}_u - \mathcal{E}_m} > 0$

Fig. 7.7 Illustration of revel repulsion due to perturbation.

orthogonal, meaning they did not "talk" to each other. The perturbation mixes the states, and makes them talk. The magnitude by which the energy of a state $|u\rangle$ changes due to interactions with *all other states* upon perturbation is $\Delta E^{(2)} \approx \sum_{m\neq u} |W_{mu}|^2/(E_u - E_m')$.

We also note the nature of the interaction. If a state E_u is interacting with states with energies E_m' lower than itself, then $\Delta E^{(2)} > 0$, the perturbation pushes the energy *up*. Similarly, interactions with states with higher energies pushes the energy of state E_u *down*. **Thus, the second-order interaction term in perturbation is *repulsive*.** Fig. 7.7 illustrates this effect schematically. This repulsive interaction is the key to understanding curvatures of energy bands and the relation between effective masses and energy bandgaps of semiconductors. Clearly if two states were non-degenerate and the strength of the perturbation is increased from zero, the energy eigenvalues repel stronger, and the levels go farther apart. Then they cannot cross each other[1].

7.5 The Hellmann–Feynman theorem

The following theorem is sometimes useful for obtaining quick estimates of the magnitude and direction of the shift in energy eigenvalues upon perturbation. Let $\hat{H}^0(\lambda)|k\rangle = E_k(\lambda)|k\rangle$ be the exactly solved problem with normalized eigenstates $|k\rangle$, where the Hamiltonian $\hat{H}^0(\lambda)$ and its resulting eigenvalues $E_k(\lambda)$ depend on a parameter λ. When we change the parameter λ in the Hamiltonian operator $\hat{H}^0(\lambda)$, how do the eigenvalues $E_k(\lambda)$ change?

Because the original eigenstates are orthonormal, we have $\langle k|k\rangle = 1$. Differentiating with respect to the parameter λ, we have $\frac{d}{d\lambda}\langle k|k\rangle = 0$.

---

[1] This goes by the name of **no level crossing theorem** of perturbation theory in quantum mechanics.

Now, applying the chain rule for differentiation,

$$\frac{d}{d\lambda}E_k(\lambda) = \frac{d}{d\lambda}\langle k|\hat{H}^0(\lambda)|k\rangle \tag{7.26}$$

$$= \langle\frac{d}{d\lambda}k|\hat{H}^0(\lambda)|k\rangle + \langle k|\frac{d}{d\lambda}\hat{H}^0(\lambda)|k\rangle + \langle k|\hat{H}^0(\lambda)|\frac{d}{d\lambda}k\rangle,$$

and because $\hat{H}^0(\lambda)|k\rangle = E_k(\lambda)|k\rangle$ and $\langle k|\hat{H}^0(\lambda) = E_k(\lambda)\langle k|$, we get

$$\frac{d}{d\lambda}E_k(\lambda) = \langle k|\frac{d}{d\lambda}\hat{H}^0(\lambda)|k\rangle + E_k(\lambda)\underbrace{[\langle\frac{d}{d\lambda}k|k\rangle + \langle k|\frac{d}{d\lambda}k\rangle]}_{\frac{d}{d\lambda}\langle k|k\rangle=0} \tag{7.27}$$

$$\implies \boxed{\frac{dE_k(\lambda)}{d\lambda} = \langle k|\frac{d\hat{H}^0(\lambda)}{d\lambda}|k\rangle.}$$

The boxed equation above is the **Hellmann–Feynman** theorem. It states that we can get the perturbation in the energy eigenvalues due to a parameter λ by finding the inner product of the derivative of the Hamiltonian operator with respect to the variable λ.

7.6 Perturbation theory example

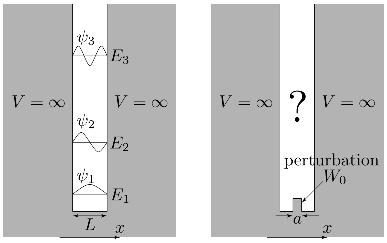


Fig. 7.8 A perturbation to the particle-in-a-box problem.

Consider the particle-in-a-box problem that was discussed in Chapter 5, Section 5.6, as shown in Fig. 7.8. Here is the question: we know the exactly solved particle in a box eigenfunctions of state $|n\rangle$: $\langle x|n\rangle = \psi_n(x) = \sqrt{\frac{2}{L}}\sin(n\frac{\pi}{L}x)$ and corresponding eigenvalues $E_n = \frac{\hbar^2}{2m_e}(n\frac{\pi}{L})^2$. Introduce a tiny perturbation potential of strength $W(x) = +W_0$ over a length $a << L$, and $W_0 << E_1$, as shown. Find the perturbed values of the eigenvalues and eigenfunctions for the states $|1\rangle$ and $|2\rangle$. We qualitatively expect the perturbation to push the particle away from the center of the well, and push energy eigenvalues higher. Let us see what perturbation theory tells us quantitatively.

Brillouin–Wigner and Rayleigh–Schrödinger Perturbation theories: Because the perturbation is $W(x) = W_0[\Theta(x - \frac{L-a}{2}) - \Theta(x - \frac{L+a}{2})]$ (This is a complicated way of writing a rectangular barrier, the meaning is very simple!) the new eigenvalues from the more general BW theory are

$$E = E_u + \langle u|W(x)|u\rangle + \sum_{m\neq u}\frac{|\langle u|W|m\rangle|^2}{E - E_u}. \tag{7.28}$$

Denoting $u = n$ as the unperturbed state, we find the first order perturbation shift:

$$W_{nn} = \langle n|W(x)|n\rangle = \frac{2W_0}{L}\int_{\frac{L-a}{2}}^{\frac{L+a}{2}} dx\,\sin^2(n\frac{\pi}{L}x) \tag{7.29}$$

$$= W_0\frac{a}{L}[1 - \cos(n\pi)\frac{\sin(n\frac{\pi}{L}a)}{(n\frac{\pi}{L}a)}].$$

Note that for even $n = 2, 4, 6, ...$, $\cos(n\pi) = +1$, and for odd $n = 1, 3, 5...$, $\cos(n\pi) = -1$. Because the perturbation extends over a small length $a << L$, we can use $\frac{\sin x}{x} \approx 1 - \frac{x^2}{6} + ...$ to get for even n states $\langle n|W(x)|n\rangle \approx W_0\frac{a}{L}\frac{(n\frac{\pi}{L}a)^2}{6} = W_0\frac{n^2\pi^2a^3}{6L^3}$. This is a *very small* change, which goes as $(\frac{a}{L})^3$ compared to the odd states for which we get $\langle n|W(x)|n\rangle \approx W_0\frac{a}{L}[2 + \frac{(n\frac{\pi}{L}a)^2}{6}] \approx 2W_0\frac{a}{L}$, which goes as $\frac{a}{L}$. Why are the odd states so strongly perturbed compared to the even states? The answer lies in the symmetry of the eigenfunctions of the unperturbed states. Because the first order perturbation $\langle n|W(x)|n\rangle$ is due to the perturbation potential $W(x)$ coupling the eigenstates $|n\rangle$ with itself, from Fig. 7.9 we see that the perturbation potential is *at the maxima* of the odd eigenfunctions (odd around $x = 0$, but even around $x = L/2$), and *at the nodes or zeroes* of the even eigenfunctions (even around $x = 0$, but odd around $x = L/2$). Because the squared integral of the even eigenfunctions is non-zero for the small interval $(\frac{L-a}{2} < x < \frac{L+a}{2})$, the even eigenfunctions still get perturbed, but by a minuscule amount. Because the odd eigenfunctions reach their maxima in the same interval, they are perturbed by a large amount.

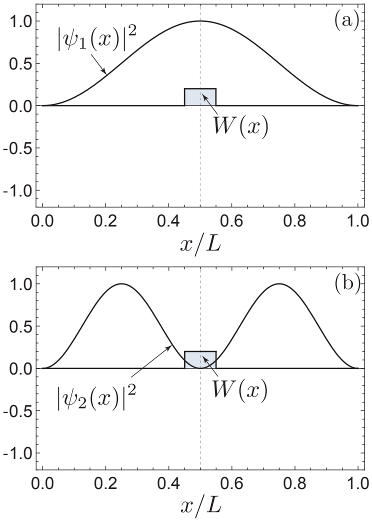


Fig. 7.9 Unperturbed eigenfunctions and probabilities. Symmetry determines if the perturbation affects certain eigenvalues, or not. (a) The $n = 1$ state will be strongly affected because its probability peaks at the location of the perturbation. (b) Because the probability of the $n = 2$ state is a minimum at the location of the perturbation, it is minimally affected.

The matrix element in the second-order eigenvalue perturbation term $\langle n|W(x)|m\rangle$ is then

$$W_{nm} = \langle n|W(x)|m\rangle = \frac{2W_0}{L}\int_{\frac{L-a}{2}}^{\frac{L+a}{2}} dx \sin(n\frac{\pi}{L}x)\sin(m\frac{\pi}{L}x). \tag{7.30}$$

We can evaluate the integral and get a most general analytical expression which is long and uninspiring. But by invoking symmetry, we can predict the fate of the eigenvalues due to the second order perturbation terms. Consider Fig. 7.10. Since the perturbation is symmetric around $x = \frac{L}{2}$, we argue that the integral is zero whenever n is odd and m is even, or if n is even and m is odd. The matrix element is non-zero only when both n, m are odd or even. This means the perturbation potential is capable of only coupling say state $|1\rangle$ with $|3\rangle$, $|5\rangle$, ..., but state $|1\rangle$ cannot couple to (or "talk to") states $|2\rangle$, $|4\rangle$, ... This is an example of a **selection rule** for coupling.

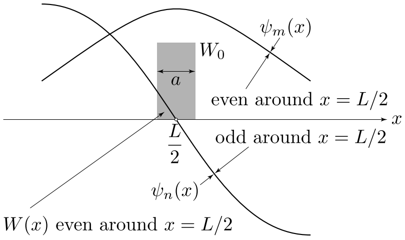


Fig. 7.10 Allowed and forbidden matrix elements obtained from the symmetry of eigenfunctions and the perturbation.

Let us find the matrix elements for the ground state $|1\rangle$. The matrix elements needed are

$$W_{m1} = \langle m|W(x)|1\rangle = 4W_0\frac{\sin(\frac{m\pi}{2})}{(m^2-1)\pi}[m\cos(\frac{\pi a}{2L})\sin(m\frac{\pi a}{2L}) - \sin(\frac{\pi a}{2L})\cos(m\frac{\pi a}{2L})]. \tag{7.31}$$

Because for even m, the term $\sin(\frac{m\pi}{2}) = 0$, state $|1\rangle$ does not couple to any other even state due to the perturbation. For all m =odd states, we make use of the small arguments of the sine and cosine above to estimate $W_{m1} \approx \frac{2}{L}W_0 a$, and write the second order perturbation terms as $\frac{|W_{m1}|^2}{E-E_m}$, as shown in detail in Fig. 7.11.

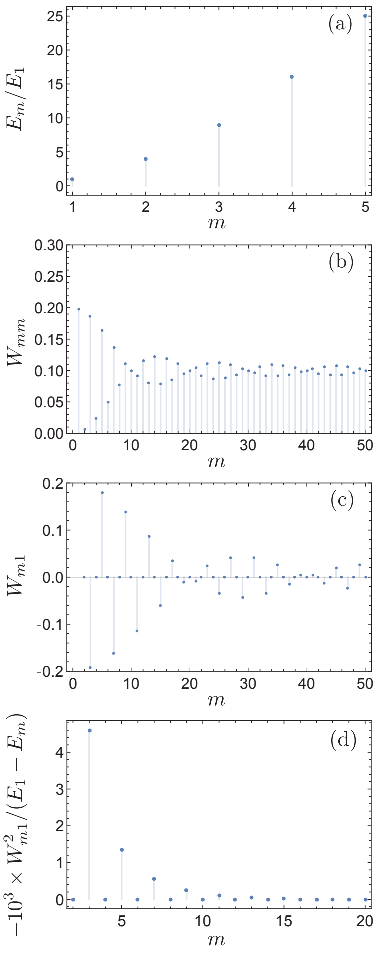


Fig. 7.11 (a) Unperturbed eigenvalues increasing as m^2. (b) Diagonal or 1st-order perturbation W_{mm} alternates between strong and weak before stabilizing for higher states at the spatially averaged perturbation $\approx W_0 \frac{a}{L}$. (c) Off-diagonal matrix elements W_{m1} of the ground state $u = 1$ with higher states m. (d) 2nd order perturbation shift decays rapidly with increasing energy separation. Note the 2nd order perturbation terms are scaled by a factor of 10^3.

Brillouin–Wigner perturbation theory: We then combine the first and second order perturbation terms to write the perturbed eigenvalue to be

$$E \approx E_1 + \frac{2a}{L} W_0 + \sum_{m=3,5,...} \frac{(\frac{2aW_0}{L})^2}{E - [\frac{\hbar^2}{2m_e}(\frac{m\pi}{L})^2 + \frac{2a}{L} W_0]}. \tag{7.32}$$

This is the BW result. To make progress, we must solve the above polynomial equation. Let us denote $E_0 = \frac{2a}{L} W_0$ If we consider only $m = 3$ in the second order sum and neglect $m = 5, 7, ...$, we get $(E - E_1 - E_0)(E - E_3 - E_0) \approx E_0^2$, which give the eigenvalues $E_1' = \frac{1}{2}[(E_1 + E_3) + 2E_0 - \sqrt{(E_3 - E_1)^2 + 4E_0^2}]$, and $E_3' = \frac{1}{2}[(E_1 + E_3) + 2E_0 + \sqrt{(E_3 - E_1)^2 + 4E_0^2}]$. If $E_0 = 0$, we recover the original eigenvalues. Since the unperturbed energy eigenvalues are non-degenerate in this problem, it is illuminating to apply the RS perturbation theory.

Rayleigh–Schrödinger perturbation theory: The RS theory from Equation 7.24 gives the perturbed ground state energy

$$E_1' \approx E_1 + E_0 + \sum_{m=3,5,...} \frac{\frac{E_0^2}{E_1}}{1 - m^2 - \frac{E_0}{E_1}} \approx E_1 + E_0 - \frac{E_0^2}{8E_1}. \tag{7.33}$$

Because m^2 increases rapidly with m and far exceeds $1 - \frac{E_0}{E_0}$, we get an acceptable estimate by retaining only the $m = 3$ term: $E_1' \approx E_1 + E_0 - \frac{E_0^2}{8E_1}$. The two perturbation terms can be now understood: the first-order "DC" shift due to the perturbation is E_0, and the second order downward energy shift $-E_0^2/8E_1$ is because the perturbation couples state $|1\rangle$ with state $|3\rangle$, the interaction is *always* replusive, and state $|3\rangle$ pushes state $|1\rangle$ energy down. The denominator $8E_1 = E_3 - E_1$ is the energy separation of the two states that are interacting due to the perturbation.

We can also find the change in the eigenfunction using Equation 7.22. To find how the ground state wavefunction has changed, we write

$$|\psi\rangle = |1\rangle + \sum_{m=3,5,...} \frac{E_0}{E_1(1 - m^2 - \frac{E_0}{E_1})} |m\rangle \tag{7.34}$$

$$\implies \psi(x) = \psi_1(x) + \sum_{m=3,5,...} \frac{E_0}{E_1(1 - m^2 - \frac{E_0}{E_1})} \psi_m(x).$$

Because the unperturbed eigenfunctions are $\psi_n(x) = \sqrt{\frac{2}{L}} \sin(n\frac{\pi}{L}x)$, we can make a plot of the new eigenfunction, and specifically its square to find the probability distribution of the ground state. Fig. 7.12 shows that the peak at the center $x = L/2$ is reduced from the unperturbed value. The perturbing potential at the center of the well pushes the ground state away from the center.

Matrix method: As the most general method of perturbation theory, we apply the matrix method to obtain the perturbed eigenvalues and eigenfunctions. Let us consider the states $m = 1, 2, 3$, and write down the matrix for the perturbed Hamiltonian $H^0 + \hat{W}$:

$$\hat{H}^0 + W = \begin{array}{c c} & \begin{array}{ccc} |1\rangle & |2\rangle & |3\rangle \end{array} \\ \begin{array}{c} \langle 1| \\ \langle 2| \\ \langle 3| \end{array} & \begin{pmatrix} E_1 + E_0 & 0 & E_0 \\ 0 & E_2 & 0 \\ E_0 & 0 & E_3 + E_0 \end{pmatrix} \end{array}, \tag{7.35}$$

and either by hand, or with a package, find the eigenvalues and eigenfunctions of this set. We did an example in Chapter 6. Now this is a very powerful technique, and matrix algorithms implemented on computers are efficient and fast. Though we are doing this example with a restricted basis of three states, it scales to larger basis reasonably well.

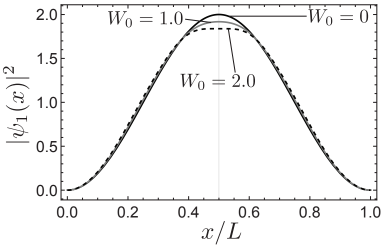


Fig. 7.12 The perturbation pushes some of the probability out from the center with increasing strength of the perturbation W_0. This is expected also from purely classical grounds.

Fig. 7.13 shows the outputs of the matrix perturbation theory with the three-set problem. The perturbed eigenvalues are plotted as a function of the perturbation strength. The perturbation strength is greatly exaggerated for this example, but the beauty of the Matrix version is that it is NOT restricted like the BW or RS versions are. The only approximation here is that we are choosing a restricted basis set.

The effects of the perturbation are evident in the movement of the eigenvalues with W_0. State $|2\rangle$ energy is unperturbed, because of symmetry arguments made above. States $|1\rangle$ and $|3\rangle$ are pushed up in energy due to the perturbation potential. However, due to the mutual repulsion between states $|3\rangle$ and $|1\rangle$, state $|3\rangle$ is pushed up, and state $|1\rangle$ is pushed down, leading to noticeable curvatures. Finally, we note that when the perturbation becomes extremely strong, state $|1\rangle$ merges with state $|2\rangle$. This is the limit of a Dirac delta potential at the center of a particle-in-a-box problem: only the even eigenvalues are allowed because the wavefunction must go to zero at the location of the Dirac delta potential. The shift in the wavefunction and probability distribution for the ground state due to perturbation calculated from the matrix method is shown in Fig. 7.13 too. These are approximate because of the small basis set, but tell the correct story.

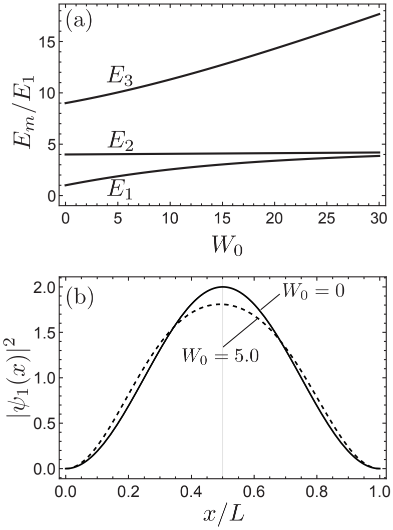


Fig. 7.13 (a) The perturbation of the three lowest eigenvalues neglecting all higher states for strong W_0, calculated by the matrix method. The effect on the 2nd state is forbidden by symmetry, and the repulsive interaction between states $m = 1$ and $m = 3$ are evident. (b) The matrix formulation also gives the expansion coefficients for the perturbed wavefunctions from which the new eigenfunctions can be created. The squared wavefunction of the ground state is shown from the matrix method as an example.

7.7 Chapter summary section

In this chapter, we learned:

- The matrix method provides the most general method to find the perturbed eigenvalues and eigenfunctions of a quantum mechanical system. The matrix method depends on the expansion principle, which guarantees that the perturbed eigenfunctions $|\psi\rangle$ can be expressed as a linear combination of the unperturbed eigenfunctions $|n\rangle$ in the form $|\psi\rangle = \sum a_n |n\rangle$. We need to be judicious in choosing the finite basis set for the matrix to capture the physical change and be numerically tractable.

- The general nature of the perturbation W is captured by its effect on pairs of unperturbed states $|n_1\rangle$ and $|n_2\rangle = \int d^3x\psi^\star_{n2}(r)W(r)\psi_{n1}(r)$. If the matrix element $\langle n_2|W|n_1\rangle = 0$, the perturbation has no effect on these eigenvalues. If the matrix element is non-zero, the new eigenfunctions are superpositions of $|n_1\rangle$ and $|n_2\rangle$, with corresponding changes in eigenvalues.
- When the unperturbed eigenvalues are non-degenerate in energy, the perturbed eigenvalues can be obtained analytically to second order by using the Rayleigh–Schrödinger version of perturbation theory, which states that for uperturbed eigenfunction $|n\rangle$ and eigenvalue E_n, the new eigenvalues are $E \approx E_n + \langle n|W|n\rangle + \sum_{n' \neq n} \frac{|\langle n'|W|n\rangle|^2}{E_n - E_{n'}}$, with the corresponding new eigenfunction $|\psi\rangle \approx |n\rangle + \sum_{n' \neq n} \frac{\langle n'|W|n\rangle}{E_n - E_{n'}}|n'\rangle$.
- The second order perturbation term is such that energy eigenvalues "repel" each other, irrespective of the sign of the perturbation. This means that the effect of perturbation is to push the higher energy level up, and the lower energy eigenvalue down. This is also called level repulsion, and ensures that there can be no crossing of energy levels due to perturbation.
- For energy levels that are degenerate, one must resort back to the matrix version to avoid unphysical blowing up of the second order Rayleigh–Schrödinger result. For small systems, the Brillouin–Wigner analytical result may be used, but it is very similar to the matrix version of perturbation theory.

Further reading

Perturbation theory is treated in most books on quantum mechanics. **Modern Quantum Mechanics** by J. J. Sakurai has an impressive list of problems to apply notions of this topic. **Quantum Mechanics** by H. Kroemer develops the topic and applies it to several problems of solid state and semiconductor physics. For a more general mathematical introduction of perturbation theories to many branches of physics, I highly recommend the classic **Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory** by C. Bender and S. Orszag.

Exercises

(7.1) **Perturbation of matrices**

Consider the general perturbation problem for a Hamiltonian matrix expressed in the unperturbed basis:

$$\begin{bmatrix} E_1 + W_{11} & W_{12} & \dots & W_{1N} \\ W_{21} & E_2 + W_{22} & \dots & W_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \dots & E_N + W_{NN} \end{bmatrix}. \tag{7.36}$$

Here, $E_1, E_2, ...E_N$ are the original unperturbed energy eigenvalues, and $W_{mn} = \int d^3r \psi_m^\star(r) W(r) \psi_n(r)$ are the matrix elements due to the perturbing potential $W(r)$.

(a) If the perturbation potential is a constant in space, $W_0(r) = W_0$, show that it only shifts the diagonal elements by $W_{nn} = W_0$, and leaves the off-diagonal elements unchanged $W_{nm} = 0$ for $n \neq m$. Show that this perturbation shifts all eigenvalues to $E_n + W_0$ for $n = 1, ..., N$ and leaves the eigenfunctions unperturbed.

(b) Consider a perturbation potential $W_1(r)$ such that all the matrix elements are equal, i.e. $W_{mn} = W_1$ for all m, n including the diagonal elements. Furthermore, assume that all the unperturbed eigenvalues are equal, and are $E_n = 0$. In this case, show that the N new eigenvalues are $(N-1)$ degenerate eigenvalues at energy 0 remain unchanged, and a single eigenvalue splits off to an energy NW_1.

(c) Finally consider a perturbation potential $W_2(r)$ such that all the off-diagonal matrix elements are equal, i.e. $W_{mn} = W_2$ for $m \neq n$, and the diagonal perturbation terms are zero, i.e. $W_{nn} = 0$. Assume that all the unperturbed eigenvalues are equal, and are $E_n = 0$. In this case, show that the N new eigenvalues are $(N-1)$ degenerate eigenvalues at energy $-W_2$, and a single eigenvalue splits off to an energy $(N-1)W_2$.

The results are shown in Fig. 7.14. Part (a) is encountered when a spatially uniform potential is applied to a doped semiconductor – all allowed eigenvalues are moved by the ”DC” value. Parts (b) and (c) are applicable when localized defect states or dopants are introduced into a semiconductor, or when crystal-fields split the valence bands. We will encounter these cases in the later chapters.

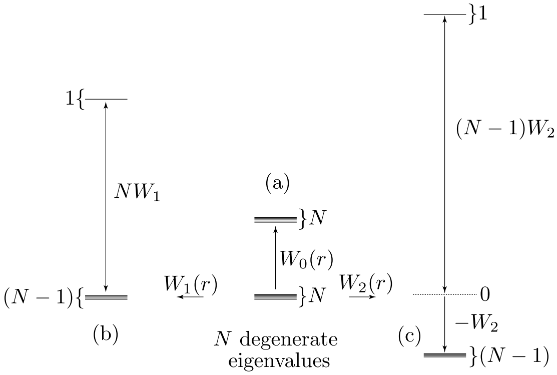


Fig. 7.14 Three types of perturbations to degenerate eigenvalues that are encountered in semiconductor physics.

(7.2) **Energy gap from a square wave potential**

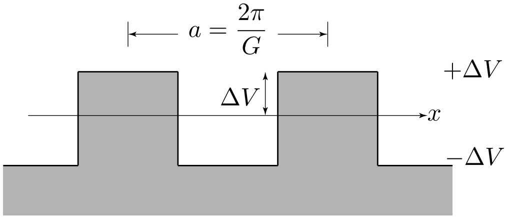


Fig. 7.15 Square-wave perturbation to the free electron.

Consider Fig. 7.15 as a periodic square wave potential in the path of a free electron moving in one dimension. The periodicity of the potential is $a = \frac{2\pi}{G}$, and the amplitude of the perturbing potential is $\Delta V = \frac{\hbar^2 G^2}{2m_e}$, where m_e is the free electron mass.

(a) Identify the set of unperturbed eigenfunctions and eigenvalues, and sketch the eigenvalue spectrum.

(b) Find the general matrix element of the perturbation potential for any two unperturbed electron eigenstates. Make the connection of the matrix element to the Fourier-representation of the perturbing potential.

(c) Outline the matrix method for solving this perturbation problem. Which set of unperturbed eigenfunctions will see the highest perturbation?

(d) Use the second order Brillouin–Wigner perturbation theory to calculate the energy bandgap due to this perturbation.

(e) Test the Rayleigh–Schrödinger (RS) approximation for the second order energy corrections. Is the RS approximation a good one for this problem?

(7.3) **Optical transitions: the quantum-confined Stark effect**

When an electric field is applied across a quantum well, the fundamental optical transitions shift in energy. This "Stark" shift of the confined levels in a quantum well can be calculated by using second-order perturbation theory. Consider the interaction between the electrons (effective mass $m^\star$) in a quantum well of width d and a DC electric field of strength F_z applied along the z (growth) axis to an electron of charge q.

(a) Explain why the perturbation to the energy of an electron is given by $W = qF_z z$. Draw a band diagram to depict the situation.

(b) Explain why the first-order shift of the energy levels, given by

$$\Delta E_n^{(1)} = \int_{-\infty}^{+\infty} dz \phi_n^\star(z) W \phi_n(z) \tag{7.37}$$

is **zero**. Use the effective mass approximation to write down the wavefunctions in an infinitely deep quantum well for this problem, which should yield $\phi_n(z)$ as the usual cosine functions.

(c) The second-order energy shift of the $n = 1$ level is given by the Rayleigh–Schrödinger approximation

$$\Delta E_{n=1}^{(2)} = \sum_{n>1} \frac{|\langle 1|W|n\rangle|^2}{E_1 - E_n}, \tag{7.38}$$

where

$$\langle 1|W|n\rangle = \int_{-\infty}^{+\infty} dz \phi_1^\star(z) W \phi_n(z). \tag{7.39}$$

Show that for an electron in a quantum well with infinitely high barriers, the Stark shift is given approximately by

$$\Delta E = -24(\frac{2}{3\pi})^6 \frac{e^2 m^\star d^4}{\hbar^2} F_z^2. \tag{7.40}$$

(d) Note that the QCSE-shift goes as $\Delta E \sim m^\star d^4 F_z^2$. Calculate and plot the shift (in meV) for a $d = 5$ nm thick AlGaN/GaN/AlGaN quantum well as a function of electric fields over a range typical in semiconductor devices. Comment on how this shifts the interband transition wavelength (red or blue?) by using same arguments for holes.

(e) The electric field F_z may be externally applied, or can arise due *internal* fields present in III-V nitride semiconductors. Assuming a polarization field of $F_z \sim 2$ MV/cm, what are the shifts in energy eigenvalues of the ground state electrons and holes?

(7.4) **Degenerate perturbation theory: application**

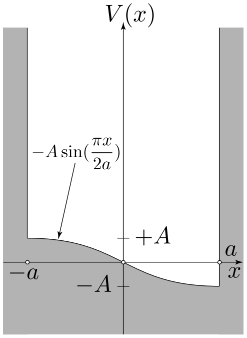


Fig. 7.16 Sinusoid perturbation to the particle in a box problem.

A particle-in a box with infinite potential barriers is perturbed by a sinusoidally deformed bottom with potential

$$V(x) = -A\sin(\pi \frac{x}{2a}) \tag{7.41}$$

within the well, and zero outside as shown in Fig. 7.16.

(a) List the unperturbed eigenvalues and corresponding eigenfunctions of the infinite quantum well with a flat bottom, when $A = 0$.

(b) Calculate the approximate energy eigenvalues for the two lowest energy eigenvalues in this potential, using a linear superposition of the two lowest states of the unperturbed well as the trial wave function. Give closed form analytic expressions for the two energies as functions of the strength of perturbation A.

(c) Assume that $A = E_1$, where E_1 is the lowest energy eigenvalue of the unperturbed well. For this specific case, indicate the energies of the two states on the graph of the potential. Give numerical values for the two energies, in eV. Make a quantitative plot of the wavefunctions for the two states.

(7.5) **A parabolic perturbation potential**

A particle-in a box with infinite potential barriers is perturbed by a sinusoidally deformed bottom with potential

$$V(x) = A[1 - (\frac{x}{a})^2] \tag{7.42}$$

within the well $-a \leq x \leq a$, and zero outside as shown in Fig. 7.17.

(a) List the unperturbed eigenvalues and corresponding eigenfunctions of the infinite quantum well with a flat bottom, when $A = 0$.

(b) Calculate the approximate energy eigenvalues for the four lowest energy eigenvalues in this potential, using a linear superposition of the two lowest states of the unperturbed well as the trial wave function. Give closed form analytic expressions for the two energies as functions of the strength of perturbation A.

(c) Assume that $A = E_1$, where E_1 is the lowest energy eigenvalue of the unperturbed well. For this specific case, indicate the energies of the four states on the graph of the potential. Give numerical values for the two energies, in eV. Make a quantitative plot of the wavefunctions for the four states.

This sort of perturbation will appear in thin semiconductor heterostructure quantum wells or transistors that have electrons in them. The curvature in the potential well will arise due to the Coulomb potential due to the presence of the charged electrons from the Poisson equation.

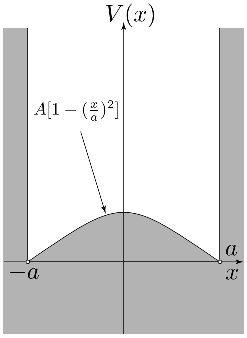


Fig. 7.17 Parabolic perturbation to the particle in a box problem.

Part II

Bands, Doping, and Heterostructures

Electrons in a Crystal Get Their Bands, Gaps, and Masses

8

8.1 The free–electron 156
8.2 Periodic perturbation 157
8.3 Bands, gaps, and effective masses 158
8.4 Non-degenerate perturbation theory 165
8.5 Glimpses of the Bloch theorem 166
8.6 Non-periodic potentials and scattering 167
8.7 Chapter summary section 169
Further reading 170
Exercises 170

The Sommerfeld and Bethe free–electron theory of electrons discussed in Chapter 5 could explain most measured properties of a metal. But there were several flaws in it. For example, though electrons are negatively charged, the Hall effect measurement of some conductors yielded a positive sign, which could not be explained by the theory. The theory disregarded the presence of atoms in the path of electrons. As a more severe flaw, even in late 1920s, the theory offered no explanation for the central *existential* question: why are some solids metals, and others insulators?

The late 1920s was when most of quantum mechanics was formalized by the likes of Schrödinger, Pauli, Heisenberg, Dirac, and their contemporaries. A rather remarkable yet bewildering logical consequence of the wave nature of electrons when they are put in a periodic potential was discovered by Felix Bloch when he was Heisenberg's student at Leipzig. Bloch showed that if the electron experiences a potential $V(x+a) = V(x)$ of periodicity a, then the electron wavefunction is of the form $\psi_k(x) = e^{ikx}u_k(x)$, where $u_k(x+a) = u_k(x)$. In this chapter, we will be able to explain why this must be the case from a perturbative route.

The goals of this chapter are to understand:

- How are the free–electron eigenfunctions and energy eigenvalues modified in a periodic potential of a crystal?
- Which free–electron states are most violently modified, and which states are left almost unaltered by the periodic potential?
- How does the periodic potential change the continuum of allowed energies by introducing **bands** of allowed energies separated by **energy gaps**?
- How do the electron energy states near the edges of allowed bands acquire an **effective mass**, and what are the concepts of electrons and **holes**?
- How do we classify the electronic properties of crystals into **metals**, **semimetals**, **semiconductors**, and **insulators**?

Fig. 8.1 Alan Wilson in 1930s explained how the number of electrons and periodic arrangement of atoms decides if a solid is a metal, a semiconductor, or an insulator during his visit in Heisenberg's research group. The idea is similar to the formation of open and closed shells for electrons in atoms periodically as the electron number increases. The formation of a crystal causes the formation of bands and gaps for electron energies; partially filled bands are conductive, similar to chemically reactive elements with open shells. If the highest band is completely filled with a significant energy gap, it is an insulator.

We discussed in Chapters 3 and 5 how the Pauli exclusion principle led to periodic repetition of chemical properties for the elements of the periodic table. To fill the allowed electron shell orbitals allowed by the Schrödinger equation, as we add electrons to the orbitals (and protons and neutrons to the nucleus) to create new elements, every now and then there are situations when a shell becomes completely filled, and the element becomes chemically inert. That is the reason behind the existence of the noble gases He, Ne, Ar, ... If we extend this concept to the *conduction* electron states in a *crystal*, it may be possible that as we increase the electron concentration, the conductivity increases, but at a certain level the conductivity decreases and the allowed "electron shells" in a crystal "close", and the crystal becomes electrically *inert*, or in other words, we have an insulating solid. This should periodically repeat. Could this analogy be the explanation for the periodic occurrence of metallic and insulating crystals as we traverse the periodic table? Alan Wilson (Fig. 8.1) explained how electrons in solids lead to metallic, semiconducting, or insulating behavior depending on their number, and the presence of the periodic potential due to the atoms in the crystal.

In Chapter 7, we developed the formalism for time-independent perturbation theory. In this chapter, we apply the theory to a free–electron perturbed by **a periodic potential**. The results we obtain will highlight most of the fundamental properties of semiconductors. These include their **energy bandstructure** $E(k)$ and opening of **bandgaps** E_g, evolution of **effective masses** $m^\star$ of various bands, work function, interactions between electron states in solids, and the role of **defects** on the interactions between electron states. In brief, this chapter will capture the essence of time-independent semiconductor physics. We start from the free–electron problem.

8.1 The free–electron

In Chapter 5 we have discussed the free–electron problem in 1D in quantum mechanics. The potential term in the Schrödinger equation is $V(x) = 0$. The eigenvectors $|k\rangle$ are such that their real-space projection yields the plane wave-function with periodic boundary conditions over a circular track of length L

$$\langle x|k\rangle = \psi_k(x) = \frac{1}{\sqrt{L}}e^{ikx}, \tag{8.1}$$

with corresponding eigenvalues

$$E_0(k) = \frac{\hbar^2 k^2}{2m_e}, \tag{8.2}$$

where $m_e \sim 9.1 \times 10^{-31}$kg is the free–electron mass. We work in a periodic-boundary condition picture, which requires $\psi(x+L) = \psi(x)$, implying that the k's are discrete, given by $k_n = (2\pi/L)n$, where n is any integer. We note immediately that a cross-matrix element of the type

$$\langle k_m|k_n\rangle = \int dx \langle k_m|x\rangle\langle x|k_n\rangle = \int_0^L dx \psi^\star_{k_m}(x)\psi_{k_n}(x)$$
$$= \frac{1}{L}\int_0^L dx e^{i2\pi(n-m)x} = \delta_{n,m} \tag{8.3}$$

is a Kronecker delta function. This is of course how it should be, since the eigenvectors states $|k_n\rangle$ and $|k_m\rangle$ are mutually orthogonal if (n, m) are different, and the states are normalized to unity. The Hamiltonian matrix (Chapter 6, Equation 6.10) is thus diagonal, with the diagonal matrix elements $\langle k|\hat{H}^0|k\rangle = E_0(k)$ given by the free–electron bandstructure in Equation 8.2. The off-diagonal elements $\langle k_m|\hat{H}^0|k_n\rangle = E_n\langle k_m|k_n\rangle$ are zero because of Equation 8.3.

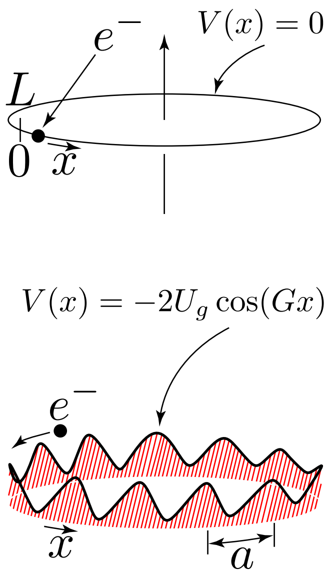


Fig. 8.2 The electron on a ring experiences a periodic potential.

8.2 Periodic perturbation

In a crystalline solid, the electron experiences a periodic potential. To model the situation for the 1D electron "on the ring" (see Fig. 8.2), let's consider the simplest periodic potential

$$V(x) = -2U_G\cos(Gx) = -U_G(e^{iGx} + e^{-iGx}), \tag{8.4}$$

where U_G is the "strength" in units of energy, and $G = 2\pi/a$, where a is the lattice-constant of the potential.

Further details of the periodic potential are shown in Fig. 8.3. The lowest energy of a *classical* particle in this potential landscape is clearly $-2U_G$, at the bottom of a valley. The Schrödinger equation for the electron is then

$$[-\frac{\hbar^2}{2m_e}\frac{\partial^2}{\partial x^2} - 2U_G\cos(Gx)]\psi(x) = E\psi(x). \tag{8.5}$$

In principle this 1D Schrödinger-equation can be solved numerically to a large degree of accuracy *directly* without perturbation theory[1]. But instead of seeking the exact solutions here, we are going to apply perturbation theory to highlight the insights it affords. The trick of perturbation theory is to consider the free–electron problem with $V(x) = 0$ (or turn the strength $U_G \to 0$) as the exactly solved problem, and consider the entire periodic potential $V(x) = -2U_G\cos(Gx) = W(x)$ as the perturbation potential.

We know the entire Hamiltonian matrix if we find the matrix elements $\langle k_2|\hat{H}^0 + W(x)|k_1\rangle = E_0(k_1)\delta_{k_1,k_2} + \langle k_2|W(x)|k_1\rangle$. The first term is the unperturbed diagonal matrix element, and the second term is due to the perturbation. The perturbation matrix element evaluates to

$$\langle k_2|W(x)|k_1\rangle = -\frac{U_G}{L}\int_0^L dx e^{i(k_1-k_2)x}(e^{iGx} + e^{-iGx})$$
$$= -U_G(\delta_{k_1-k_2,-G} + \delta_{k_1-k_2,+G}). \tag{8.6}$$

---

[1] This differential equation is known as the Mathieu equation, first introduced by Leonard Mathieu in 1868. The solutions, called Mathieu functions explain classical vibrational modes in elliptical drumheads.

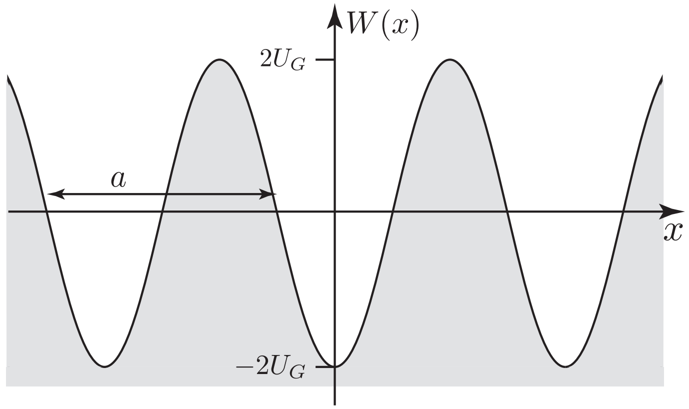


Fig. 8.3 A periodic potential $W(x) = -2U_G \cos(Gx)$ acts as a perturbation to the free–electron.

The Kronecker-delta implies that the perturbation only couples states $|k_1\rangle$ and $|k_2\rangle$ if their wavevector difference is $k_1 - k_2 = \pm G$, the reciprocal lattice vector of the perturbing potential[2]. Recall from Chapter 7 that we can find the perturbed eigenvalues by the matrix method, which works *both* for degenerate and non-degenerate states. But if we were to consider all the $|k\rangle$ states, the matrix would be ∞-dimensional. So we should choose a restricted set for identifying the eigenvalues.

8.3 Bands, gaps, and effective masses

It is clear from Equation 8.6 that a state $|k\rangle$ will interact due to the periodic perturbation with only two other states $|k+G\rangle$ and $|k-G\rangle$ directly. This will require us to solve a 3×3 Hamiltonian for the three states $|k-G\rangle$, $|k\rangle$, and $|k+G\rangle$. But also recall in Chapter 7 the result of non-degenerate perturbation theory told us that the changes in eigenvalues for states widely separated in energy goes as $|W_{12}|^2/(E_1 - E_2)$. So the states that interact *most strongly* due to the perturbation must be close (or degenerate) in energy, but their wavevectors should still follow $k_1 - k_2 = \pm G$. Clearly, two such states are $|+G/2\rangle$ and $|-G/2\rangle$. This is illustrated in Fig. 8.4. To locate states that have non-zero matrix elements, one has to imagine sliding the double-headed arrow of length G along the k-axis. Two situations are shown, one when the unperturbed states are degenerate, and one when they are not. Also remember the repulsive nature of the interaction: in Fig. 8.4 we expect state $|k_1\rangle$ to be pushed *down*, and state $|k_2\rangle$ to be pushed *up* due to their mutual interaction.

The unperturbed eigenvalue of the two degenerate states is $E_0(G/2) = \hbar^2 G^2/8m_e = F$. Clearly this is a case for the application of *degenerate* perturbation theory[3]. The problem is rather simple, since the Hamil-

[2] This condition appears in crystallography as **Bragg's law**. Any wave of wavevector $\mathbf{k}$ when interacting with a crystal of reciprocal lattice vector $\mathbf{G}$ will scatter into states $\mathbf{k} \pm \mathbf{G}$. The strongest scattering is for states $\mathbf{k} = \pm\frac{\mathbf{G}}{2}$. These states are scattered the most, and into each other.

[3] We will see later that the Brillouin–Wigner (BW) non-degenerate perturbation theory also can give the same result.

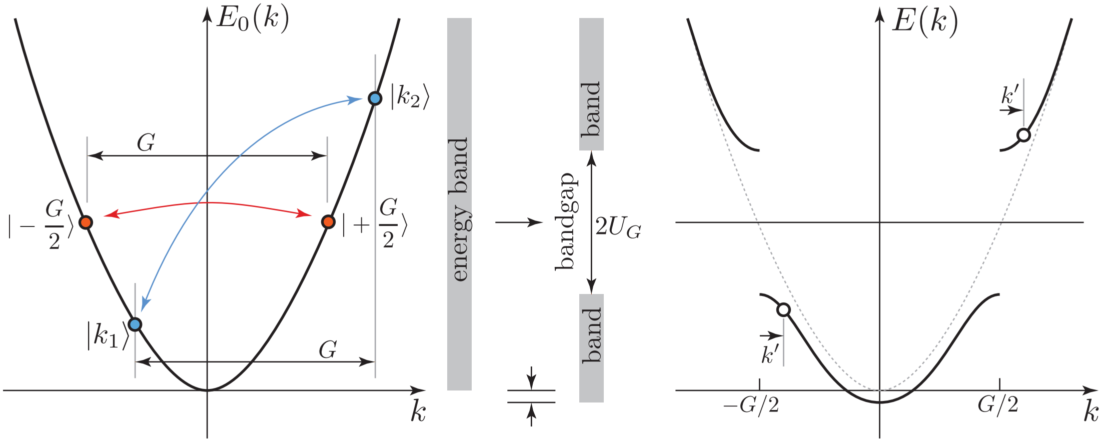


Fig. 8.4 Bandgap opening in the energy spectrum of a free–electron upon perturbation by a periodic potential.

tonian is a 2×2 matrix:

$$\hat{H}^0 + W = \begin{array}{c} \\ \langle +\frac{G}{2}| \\ \langle -\frac{G}{2}| \end{array} \begin{array}{c} \begin{array}{cc} |+\frac{G}{2}\rangle & |-\frac{G}{2}\rangle \end{array} \\ \begin{pmatrix} F & -U_G \\ -U_G & F \end{pmatrix} \end{array}, \tag{8.7}$$

where we write out the ket and bra states explicitly to highlight where the matrix elements come from. The eigenvalues of this matrix are obtained by solving the determinant of the matrix: $(F-E)^2 - U_G^2 = 0$, which yields $E_\pm = F \pm U_G$. This implies the degenerate unperturbed states $E_0(+G/2) = E_0(-G/2) = F$ have now been split to two energies E_+ and E_- with the difference $E_+ - E_- = 2U_G$ by the periodic perturbation. This is the opening of a *bandgap* in the allowed energies for the electron, and is highlighted in Fig. 8.4.

Fig. 8.5 Rudolph Peierls while working with Heisenberg was the first to produce the famous energy dispersion of a free–electron perturbed by a periodic potential. His plot is reproduced in Fig. 8.6.

We note here that the general eigenvalues of the 2×2 Hamiltonian matrix

$$\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \tag{8.8}$$

are

$$E_\pm = \frac{H_{11}+H_{22}}{2} \pm \sqrt{(\frac{H_{11}-H_{22}}{2})^2 + |H_{12}|^2}, \tag{8.9}$$

and the corresponding eigenvectors are

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}_\pm = \begin{bmatrix} \frac{H_{12}}{\sqrt{|H_{12}|^2+(E_\pm - H_{11})^2}} \\ \frac{E_\pm - H_{11}}{\sqrt{|H_{12}|^2+(E_\pm - H_{11})^2}} \end{bmatrix}. \tag{8.10}$$

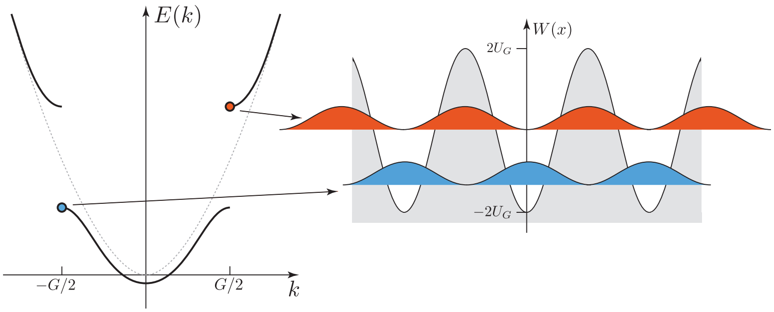


Fig. 8.7 Probability pileups of band-edge states.

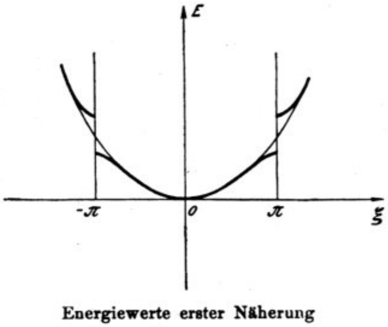


Fig. 8.6 Rudy Peierls' iconic plot of the energy dispersion of a free–electron in a periodic potential from his 1930 paper.

Since we expand the perturbed states as $|\psi\rangle = \sum_n a_n |n\rangle$, for the degenerate states, we obtain the perturbed eigenvectors as

$$|\pm\rangle = a_{1\pm}|+\frac{G}{2}\rangle + a_{2\pm}|-\frac{G}{2}\rangle. \tag{8.11}$$

For the degenerate states we get $a_{1+} = -1/\sqrt{2}$ and $a_{2+} = +1/\sqrt{2}$, and $a_{1-} = -1/\sqrt{2}$ and $a_{2-} = -1/\sqrt{2}$. The identification of the coefficients helps us convert the perturbed eigenvectors into the eigenfunctions

$$\langle x|+\rangle = \psi_+(x) = (-\frac{1}{\sqrt{2}})\cdot(\frac{e^{i\frac{G}{2}x}}{\sqrt{L}}) + (+\frac{1}{\sqrt{2}})\cdot(\frac{e^{-i\frac{G}{2}x}}{\sqrt{L}}) = -i\sqrt{\frac{2}{L}}\sin(\frac{G}{2}x), \tag{8.12}$$

and

$$\langle x|-\rangle = \psi_-(x) = (-\frac{1}{\sqrt{2}})\cdot(\frac{e^{i\frac{G}{2}x}}{\sqrt{L}}) + (-\frac{1}{\sqrt{2}})\cdot(\frac{e^{-i\frac{G}{2}x}}{\sqrt{L}}) = -\sqrt{\frac{2}{L}}\cos(\frac{G}{2}x). \tag{8.13}$$

This is illustrated in Fig. 8.7. Note now the properties of $|\psi_+(x)|^2 = (2/L)\sin^2(Gx/2)$ and $|\psi_-(x)|^2 = (2/L)\cos^2(Gx/2)$. The probability densities for the higher energy states $E_+ = F + U_G$ go as $\sin^2(Gx/2)$, meaning they peak at the highest points of the perturbing potential. The high potential energy is responsible for the high net energy of these states. Similarly, the lower energy states $E_- = F - U_G$ pile up in the valleys, and consequently have lower energies. Note that due to the perturbation, the new eigenfunctions of the degenerate states *no longer* have a uniform probability distribution in space. For the unperturbed free–electron problem, since $\psi_k(x) = \frac{1}{\sqrt{L}}e^{ikx}$, the probability density $|\psi_k(x)|^2 = \frac{1}{L}$ was uniform everywhere. The periodic perturbation has modified this distribution most significantly for the states with $k \approx \pm\frac{G}{2}$.

But what about states that are not degenerate? Let's look at the states $|k_2\rangle = |\frac{G}{2} + k'\rangle$ and $|k_1\rangle = |-\frac{G}{2} + k'\rangle$, for example those shown in Fig. 8.4. By tuning the magnitude of k', we can move as close to the $\pm G/2$ states as possible. The perturbed Hamiltonian is

$$\hat{H}^0 + W = \begin{array}{c|cc} & |+\frac{G}{2}+k'\rangle & |-\frac{G}{2}+k'\rangle \\ \hline \langle +\frac{G}{2}+k'| & E_0(+\frac{G}{2}+k') & -U_G \\ \langle -\frac{G}{2}+k'| & -U_G & E_0(-\frac{G}{2}+k') \end{array}, \tag{8.14}$$

where we write the diagonal unperturbed eigenvalues as

$$E_0(\pm\frac{G}{2}+k') = \underbrace{\frac{\hbar^2 G^2}{8m_e}}_{F} + \underbrace{\frac{\hbar^2 k'^2}{2m_e}}_{\mathcal{E}(k')} \pm \underbrace{\frac{\hbar^2 G k'}{2m_e}}_{2\sqrt{F\mathcal{E}(k')}} = F + \mathcal{E}(k') \pm 2\sqrt{F\mathcal{E}(k')}. \tag{8.15}$$

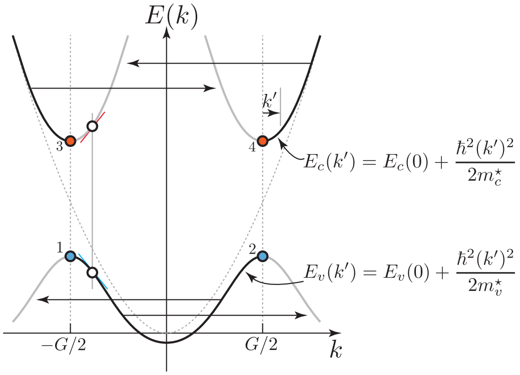


Fig. 8.8 Repeated zone scheme for energy bands in crystals, band edges, effective masses, and group velocities of allowed energy states. The group velocity $v_g = \frac{1}{\hbar}\frac{dE(k)}{dk}$ is proportional to the slope of the bands. The solid gray lines show that by translating any allowed band by wavevector $\pm G$, we can bring them back into the interval $-\frac{G}{2} \leq k \leq +\frac{G}{2}$. Typically bands are plotted only in this interval, called the first Brillouin zone. Before the opening of the bandgaps, for the free–electron, the two states with $k = \pm\frac{G}{2}$ were distinct states, one was moving to the right, and the other to the left. But after opening of the bandgaps, the states at $k = \pm\frac{G}{2}$ are actually the **same state** within each band!

The eigenvalues then are obtained from Equation 8.9 as

$$E_\pm(k') = F + \mathcal{E}(k') \pm \sqrt{4F\mathcal{E}(k') + U_G^2} \approx F + \mathcal{E}(k') \pm U_G(1 + \frac{2F\mathcal{E}(k')}{U_G^2}), \tag{8.16}$$

where we have expanded the square root term using $(1+x)^n \approx 1 + nx + \ldots$ for $x << 1$ assuming $4F\mathcal{E}(k')/U_G^2 << 1$. The energy dispersion then becomes

$$E_\pm(k') \approx (F \pm U_G) + (1 \pm \frac{2F}{U_G})\frac{\hbar^2 k'^2}{2m_e}, \tag{8.17}$$

from where we choose the + sign as a "conduction" band with lowest energy $E_c(0) = F + U_G$, and the $-$ sign as the "valence" band with highest energy $E_v(0) = F - U_G$, and bandgap $E_g = E_c(0) - E_v(0) = 2U_G$. We rewrite the energy dispersions as

$$\boxed{\begin{array}{rcl} E_c(k') & \approx & E_c(0) + \frac{\hbar^2 k'^2}{2m_c^\star} \\ E_v(k') & \approx & E_v(0) - \frac{\hbar^2 k'^2}{2m_v^\star}, \end{array}} \tag{8.18}$$

where the conduction band-edge *effective* mass and valence band-edge *effective* masses are, respectively,

$$\boxed{\begin{array}{l} m_c^\star = \frac{m_e}{1+\frac{2F}{U_G}}, \text{ and} \\ m_v^\star = -(\frac{m_e}{1-\frac{2F}{U_G}}). \end{array}} \tag{8.19}$$

These energy dispersions and **effective masses** are shown in Fig. 8.8. Note that the velocities of states 1 and 2 are zero; and since we started with two states before opening of the gap, there must be only two after opening the gap too. So state $1 = 2$, and state $3 = 4$. This is only true at the band edges. One can imagine that the $k = \frac{G}{2}$ point is identical to the $k = -\frac{G}{2}$ point; this is explored further in Exercise 8.8. We note immediately that the *effective* mass of carriers at the band-edges is different from the mass of the free–electron. The conduction

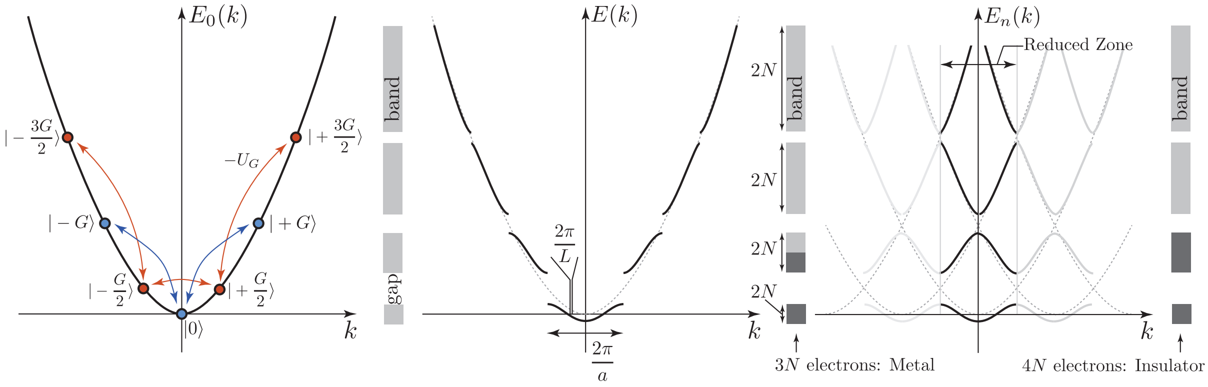

$E_0(k)$
$|-\frac{3G}{2}\rangle$
$|+\frac{3G}{2}\rangle$
$-U_G$
$|-G\rangle$
$|+G\rangle$
$|-\frac{G}{2}\rangle$
$|+\frac{G}{2}\rangle$
$|0\rangle$
k
band
gap
$E(k)$
$\frac{2\pi}{L}$
$\frac{2\pi}{a}$
k
$2N$
$2N$
$2N$
$2N$
band
$E_n(k)$
Reduced Zone
k
band
$3N$ electrons: Metal
$4N$ electrons: Insulator

Fig. 8.9 Indirect coupling via intermediate states. Each coupling has a strength $-U_G$.

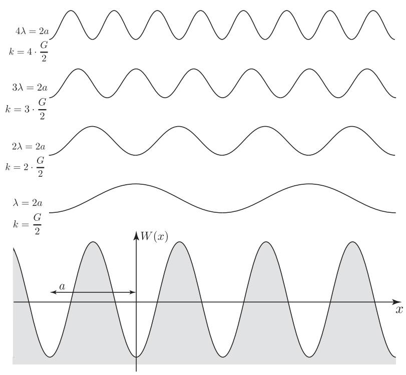

$4\lambda = 2a$
$k = 4 \cdot \frac{G}{2}$
$3\lambda = 2a$
$k = 3 \cdot \frac{G}{2}$
$2\lambda = 2a$
$k = 2 \cdot \frac{G}{2}$
$\lambda = 2a$
$k = \frac{G}{2}$
$W(x)$
a
x

Fig. 8.10 Energy gaps open at $k = n \cdot \frac{G}{2}$, which is identical to the conditions when the electron wavelengths become resonant with twice the lattice constant, i.e. $n\lambda = 2a$. This is analogous to Bragg's law of diffraction for a wave incident at an angle θ on a crystal of period d, when the law reads $n\lambda = 2d \sin\theta$.

band edge effective mass is *lower* than the free–electron mass; the electron moves as if it is lighter. If we assume that $U_G << F$, we can neglect the 1 in the denominator, and we get the interesting result that $m_c^\star \sim (U_G/2F)m_0$, that is, **the effective mass is proportional to the energy bandgap**. We will see later in Chapter 12 in the $\mathbf{k} \cdot \mathbf{p}$ theory that for most semiconductors, this is an excellent rule of thumb for the conduction band effective mass.

The valence band effective mass under the same approximation is $m_v^\star \sim \frac{U_G}{2F} m_e$. Note that we adopt the convention that the hole effective mass is *positive* in sign, but the bandstructure curves *downwards* in Fig. 8.8, so the band curvature is negative. Physically, it means that the electron in the valence band moves in the opposite direction to an electron in the conduction band in the same $|k\rangle$ state. This is clear from the **group velocity** $v_g = \frac{1}{\hbar}\frac{dE(k)}{dk}$: the slopes of the states are opposite in sign.

Are there k-states other than $|\pm G/2\rangle$ at which energy gaps develop due to the perturbation? Let's examine the states $|\pm G\rangle$, with unperturbed energy $4F$. Clearly, $k_2 - k_1 = 2G$, so there is no *direct* interaction between the states. But an *indirect* interaction of the form $|-G\rangle \leftrightarrow |0\rangle \leftrightarrow |+G\rangle$ is possible. This is illustrated in Fig. 8.9. The eigenvalues for such an interaction are found by diagonalizing the 3×3 perturbation Hamiltonian

$$\hat{H}^0 + W = \begin{array}{cccc} & |-G\rangle & 0 & |+G\rangle \\ \langle -G| & 4F & -U_G & 0 \\ \langle 0| & -U_G & 0 & -U_G \\ \langle +G| & 0 & -U_G & 4F \end{array}, \tag{8.20}$$

which yields the perturbed eigenvalues $4F, 2F \pm \sqrt{4F^2 + 2U_G^2}$. If the perturbation potential is weak, i. e., $U_G << F$, then we can expand the square root to get the three eigenvalues $4F, 4F + U_G^2/2F, -U_G^2/2F$. We note that there indeed is a splitting of the $|\pm G\rangle$ states, with an energy bandgap $U_G^2/2F$. Similarly, gaps will appear at $\pm nG/2$, due to indirect interactions[4] $|-nG/2\rangle \leftrightarrow |-(n/2+1)G\rangle \ldots \leftrightarrow |+nG/2\rangle$, with a bandgap that scales as $(U_G)^m$. For example, the indirect interaction $|-3G/2\rangle \leftrightarrow |-G/2\rangle \leftrightarrow |+G/2\rangle \leftrightarrow |+3G/2\rangle$ is depicted schematically in Fig. 8.9. Fig. 8.10 shows that the wavevectors at which gaps open are the same as Bragg's condition.

We also note that the intermediate state $|k=0\rangle$ which had a zero unperturbed energy now has been pushed down, and has a negative energy of $-U_G^2/2F$. Thus the ground state energy of the electron is now *negative*, implying it is energetically favorable that the electron be in this state. This idea develops into the concept of a **work-function** of a solid: it takes energy to kick out an electron from the ground state into the free–electron state, which has a minimum of zero. The work-function is schematically illustrated in Fig. 8.4.

Fig. 8.9 indicates several foundational concepts of electronic states in 1D crystals. These concepts carry over to crystals in higher dimensions, in 2D or 3D.[5] Because of the opening of gaps at $k = n\frac{G}{2}$, the energy bands are separated by energy gaps. Because the width of each band in the k-space is $G = \frac{2\pi}{a}$, and the separation between two allowed states is $\frac{2\pi}{L}$ where L is the macroscopic length, **the number of states in each band is $N = L/a$, equal to the number of unit cells in the crystal**. Since each allowed state can hold two electrons of opposite spin, **each band can hold 2N electrons**. We can plot all energies within $-\frac{G}{2} \leq k \leq +\frac{G}{2}$ and index them as $E_n(k)$ where n indicates a reciprocal lattice vector. Because of the opening of the bandgaps, when an electron in an allowed band moves in response to $F = \hbar\frac{dk}{dt}$, it cannot jump to the higher band when its $k \to \pm\frac{G}{2}$ under moderate forces[6]. In that case, the electron remains in the same band, but enters it from $k = -\frac{G}{2}$. This is because $k = +\frac{G}{2}$ and $k = -\frac{G}{2}$ are actually the same point in the reduced zone scheme (see Exercise 8.8).

If a band is completely filled, it is clear that a significant amount of currents are carried by individual quantum states as $v_g = \frac{1}{\hbar}\frac{dE(k)}{dk}$ is non-zero for all states except at the band edges. If these states are occupied by electrons, then there is a substantial quantum current $J_R = q\frac{g_s}{L}\sum_{+k} v_g(k)f(k)$ flowing to the right. But if a band is completely filled, this right-going current is exactly balanced by a left-going current. Indeed, $J_{filled} = q\frac{g_s}{L}\sum_k v_g(k)f(k) = 0$ because the group velocity for every $+k$ state is exactly equal and opposite to that of state $-k$. Therefore, we obtain an important quantum-mechanical result[7]:

Filled bands carry no net current.

This implies that for a 1D crystal to be a metal, each atom must have an odd number of electrons, and for it to be an insulator, it must have an even number of electrons, as indicated in Fig. 8.9. Stated in

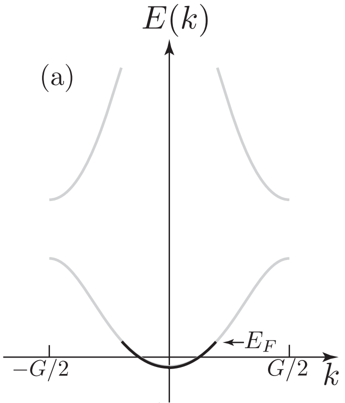
---

[4] Here n is an integer.

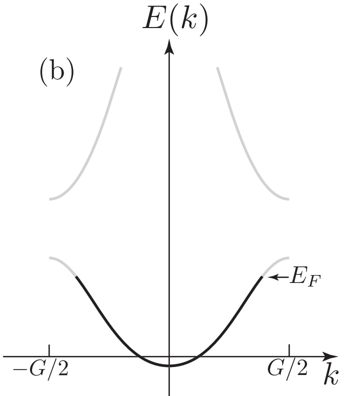
[5] The momentum of the free–electron was $\hbar k$ in the absence of the periodic perturbation potential. What is the meaning of $\hbar k$ in the bandstructure of Fig. 8.9 after the opening of the bandgaps? $\hbar k$ is now referred to as the **crystal momentum**. The rate of change of the crystal momentum is equal to an external force $F = \frac{d(\hbar k)}{dt}$. The rate of change of the electron momentum is due to both the external force, and the internal periodic potential. Because of the bandstructure $E(k)$, the effect of the crystal and the "momentum" it imparts to the electron is automatically accounted for in this picture. $\hbar k$ is *not* the momentum of the electron anymore. We will have much more to say about the crystal momentum in Chapter 9.

[6] Under very high forces, an electron can transition from one band to another through the energy gap. This is called tunneling, and will be explored in Chapter 18.

[7] To be more precise, a filled band carries no net longitudinal current. It can still carry a circulating current that flows on surfaces or edges: these (relatively small) currents are due to a Berry phase which we have not discussed.



n-type



p-type

Fig. 8.11 The band theory developed by Peierls for the first time offered an explanation for a positive measured Hall coefficient. Though the particles conducting the electric current are all electrons, the Hall effect shows a positive sign for some metals and doped semiconductors. This is because the Fermi level is near the top of a band, the effective mass is negative, and the transport can be thought of as due to the superposition of a filled band with zero conductivity, with positively charged holes at the top of the band.

a different language, if the Fermi level E_F is inside an allowed energy band, it is a metal. If the Fermi level is at the top of a band or in a gap, it is an insulator, or a semiconductor, depending on the situation.

Given this form of filling of electron band states by the Pauli exclusion principle, it is clear that as the number of electrons per atom increases, the crystal will periodically cycle between metallic and insulating phases. The insulating crystals are called semiconductors if the gap nearest the Fermi level is not too large ($\sim 1 - 5$ eV) and if the Fermi level can be controlled by techniques such as field effect, doping, or optical excitation. Though we have derived this result for a 1D crystal, it holds true for 3D: the 3D crystal of silicon has a bandgap of 1 eV, and GaN has a bandgap of 3.4 eV. Both are semiconductors. An unsolved mystery that the Sommerfeld free–electron theory of metals could not explain was that some metals and conductors experimentally showed a positive Hall coefficient. Fig. 8.11 explains this mystery. When the Fermi level E_F is at an energy where the band has a positive (upwards) curvature as in Fig. 8.11 (a), the energy dispersion resembles that of a free–electron, and the Hall-sign is negative, the charge of the electron. But when the Fermi level is at an energy where the curvature of the band is negative as in as in Fig. 8.11 (b), the bandstructure is flipped upside down compared to that of a free–electron. An electron moving to higher k values slows down, as the group velocity $v_g = \frac{1}{\hbar}\frac{\partial E(k)}{\partial k}$ decreases, reaching zero at the top of the band. The current carried by these states can be considered to be entirely due to positively charged empty states, or **holes** at the top of the band, plus a filled band with negatively charged electrons.

Since the current in a filled band is zero, the entire current is due to positively charged holes in this band, and also shows as a positive sign in the Hall coefficient. In the language of quantum currents, we can write $J = (-q)\frac{g_s}{L}\sum_k v_g(k) f(k) = (-q)\frac{g_s}{L}\sum_k v_g(k)[f(k) + 1 - 1] = (-q)\frac{g_s}{L}\sum_k v_g(k) + (+q)\frac{g_s}{L}\sum_k v_g(k)[1 - f(k)]$. Here $q = +1.6 \times 10^{-19}$ Coulomb is the magnitude of the electron charge. Since $\sum_k v_g(k) = 0$ over a full band, the quantum current in the 1D band is

$$\boxed{J = \underbrace{(+q)}_{\text{hole charge}} \frac{g_s}{L} \sum_k v_g(k) \underbrace{[1 - f(k)]}_{\text{hole occupation probability}}.} \tag{8.21}$$

Now $1 - f(k)$ is the probability that state k is **not filled by an electron**, or in other words, it is the probability that it is **filled by a hole**. Equation 8.21 shows that the current flow is due to a particle of charge $+q = +1.6 \times 10^{-19}$ Coulomb, of sign *opposite* to that of the electron. This is the reason the **hole behaves as a positively charged particle** in electrical conductivity and Hall effect. For example, because like charges repel, to remove or deplete carriers from an n-type conductor capacitively by field effect, we will need to apply a negative voltage. But a positive voltage will be necessary to deplete holes[8]. Though the result above is derived for the 1D problem, it carriers over with no changes to higher dimensions.

[8] Another way to look at this strange behavior is: by *adding* electrons, we *reduce* the conductivity, since we will fill the vacant states at the top off the band and close the p-type conductor, turning it into an insulator!

The explanation of the positive Hall coefficient by Rudy Peierls (Fig. 8.5) in this manner in 1930 was a triumph of the quantum theory of solids. The p-type case is encountered in some metals, and in heavily acceptor-doped semiconductors. The p-type conductor is analogous to atomic states that have a few outermost electronic shells empty and most shells full such as oxygen and nitrogen. The n-type conductor is analogous to atomic states that have a few outermost shells filled, and most shells empty, such as sodium and magnesium. Similarities in the optical spectra of ”n-type” and ”p-type” atoms mirror themselves in similarities in the conductivity of n-type and p-type conductors.

8.4 Non-degenerate perturbation theory

A number of results that we obtained by matrix methods in the previous section may be obtained using analytical non-degenerate perturbation theory. Recall that non-degenerate perturbation theory in Chapter 7, Equation 7.23 provided the Brillouin–Wigner (BW) result

$$E \approx E_u + \langle u|W|u\rangle + \sum_{m\neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E - E'_m}, \tag{8.22}$$

with the Rayleigh–Schrödinger result obtained by simply replacing $E \to E_u$ on the right side. Let us investigate whether we can apply it to the electron in a periodic potential problem.

If we apply the BW theory to the states $|u\rangle = |\pm G/2\rangle$, we identify $E_u = F$, $\langle u|\hat{W}|u\rangle = 0$, and $\langle -G/2|\hat{W}|+G/2\rangle = -U_G$, the sum in the RHS of Equation 8.22 has just one term, and we get

$$E \approx F + \frac{U_G^2}{E - F} \implies E \approx F \pm U_G, \tag{8.23}$$

which actually yields the same result as obtained by degenerate perturbation theory in the last section. The two degenerate states are split, with a gap of $2U_G$. This is an advantage of the BW theory: it works even for degenerate states, though it is typically classified under non-degenerate perturbation theory. Note that we had to solve the same quadratic equation as the 2×2 matrix in the degenerate theory. They are the same thing. The disadvantage of the BW theory is that it requires us to solve for the roots of a polynomial equation even in the non-degenerate situation, the payoff is the results are guaranteed to be more accurate.

Clearly the RS theory result

$$E \approx E_u + \langle u|W|u\rangle + \sum_{m\neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E_u - E'_m} \tag{8.24}$$

cannot be applied to degenerate states, since the denominator in the sum on the RHS will become zero. But it is well-suited for non-degenerate states. For example, if we ask the question how is state

$|0\rangle$ perturbed by its interaction with states $|-G\rangle$ and $|+G\rangle$, we get

$$E \approx 0 + 0 + \frac{U_G^2}{0 - 4F} + \frac{U_G^2}{0 - 4F} = -\frac{U_G^2}{2F}, \tag{8.25}$$

which is the approximate result we had obtained by diagonalizing the perturbation matrix in Equation 8.20. For small perturbations, this result is a good approximation. But if U_G increases, it is easy to see that the minimum energy $-U_G^2/2F$ can become *lower* than the classically minimum energy allowed in the system, which is $-2U_G$. This should be clear from Fig. 8.3. The minimum energy allowed for the electron should be *larger* than $-2U_G$ because of quantum confinement, implying $U_G << 4F$ for perturbation theory to be applicable.

Application of the BW theory removes this restriction, since it requires the solution of

$$E \approx 0 + 0 + \frac{U_G^2}{E - 4F} + \frac{U_G^2}{E - 4F} = -\frac{2U_G^2}{E - 4F}, \tag{8.26}$$

which yields the same result as the non-degenerate matrix method for state $|0\rangle$: $E \approx 2F - \sqrt{4F^2 + 2U_G^2}$. This root is clearly always greater than $-2U_G$, since it asymptotically approaches $-\sqrt{2}U_G$ when U_G is large. But note that too large a U_G compared to F makes the "perturbative" treatment not valid in the first place, highlighting the fact that the BW result only to the second order is also an approximation.

8.5 Glimpses of the Bloch theorem

Consider the free–electron state $|k\rangle$ with real-space projection $\psi_k(x) = \langle x|k\rangle = e^{ikx}/\sqrt{L}$. Due to the periodic potential $W(x) = -2U_G\cos(Gx)$, this state couples to the states $|k+G\rangle$ and $|k-G\rangle$. What is the perturbed *wavefunction* in real space?

From Equation 7.22 in Chapter 7, we write the perturbed state vector $|k'\rangle$ as

$$|k'\rangle \approx |k\rangle + \frac{\langle k+G|W|k\rangle}{E(k) - E(k+G)}|k+G\rangle + \frac{\langle k-G|W|k\rangle}{E(k) - E(k-G)}|k-G\rangle \tag{8.27}$$

where we have used the Rayleigh–Schrödinger version. The matrix elements are $-U_G$; projecting the perturbed state on $\langle x|$ we get the perturbed wavefunction to be

$$\psi_{k'}(x) = \langle x|k'\rangle \approx \frac{e^{ikx}}{\sqrt{L}} - \frac{U_G}{E(k) - E(k+G)}\frac{e^{i(k+G)x}}{\sqrt{L}} - \frac{U_G}{E(k) - E(k-G)}\frac{e^{i(k-G)x}}{\sqrt{L}} \tag{8.28}$$

from where we split off e^{ikx} to write the wavefunction as

$$\psi_{k'}(x) \approx e^{ikx} \cdot \underbrace{\left[\frac{1}{\sqrt{L}} - \left(\frac{U_G}{E(k) - E(k+G)}\right)\frac{e^{iGx}}{\sqrt{L}} - \left(\frac{U_G}{E(k) - E(k-G)}\right)\frac{e^{-iGx}}{\sqrt{L}}\right]}_{u_k(x)}. \tag{8.29}$$

Note that the wavefunction is of the form $e^{ikx}u_k(x)$, where the function $u_k(x)$ has the property $u_k(x+a) = u_k(x)$, because $e^{\pm iGa} = 1$. This is really the statement of the Bloch theorem: the eigenfunctions for an electron in the presence of a periodic potential can be written in the form $\psi_k(x) = e^{ikx}u_k(x)$, where $u_k(x+a) = u_k(x) = \sum_G c_G e^{iGx}$ (c_G's are constants) has the same periodicity as the potential.

$W(x) = -2[U_{G_1}\cos(G_1 x) + U_{G_2}\cos(G_2 x) + ...]$ represents a more complicated periodic potential that will lead to more couplings, and create more terms in $u_k(x)$. But the form of the Bloch decomposition of the wavefunction in Equation 8.29 will still remain true. We call this a "glimpse" of the Bloch theorem because of the "$\approx$" sign in Equation 8.29; in the next chapter this sign will be rigorously turned into an equality. The Bloch theorem is a *non-perturbative* result: it does not depend on the strength of the periodic potential. But of course we just saw it naturally *emerge* as a result from perturbation theory.

A natural consequence of the discussion above is what is called the **nearly free–electron model** of the electron bandstructure, where the periodic potential $V(x) = 0$ is turned off, but the lattice periodicity remains. In such a case, the wavefunctions must still be Bloch functions $\psi(x) = e^{ikx}\sum_G c_G e^{iGx}$, but since the Schrödinger equation is $-\frac{\hbar^2}{2m_e}\frac{\partial^2}{\partial x^2}\psi(x) = E\psi(x)$, by direct substitution we obtain the equation relating coefficients as $\sum_G c_G e^{iGx}\frac{\hbar^2(k+G)^2}{2m_e} = E\sum_G c_G e^{iGx}$. By equating coefficients, we obtain the **nearly free–electron (NFE) bandstructure**

$$\boxed{E_n(k) = \frac{\hbar^2}{2m_e}(k-nG)^2 \text{ where } n = 0, \pm 1, \pm 2, ...} \tag{8.30}$$

As shown in Fig. 8.12, the NFE bandstructure consists of copies of the free–electron bandstructure translated by nG in the k-axis, where $G = \frac{2\pi}{a}$ and $n = 0, \pm 1, \pm 2, ...$ This is a direct consequence of periodicity of the lattice. The NFE bandstructure is simple to calculate, but has profound consequences in later chapters as we step into higher dimensions.

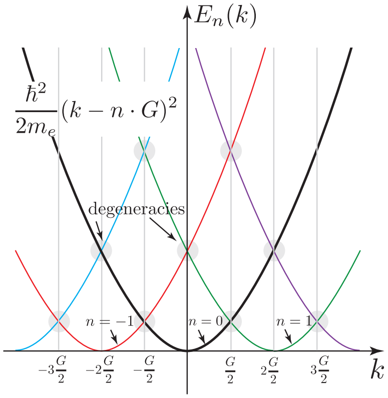


Fig. 8.12 The nearly free–electron bandstructure. The model is where there is a lattice, but no crystal potential. Bloch theorem ensures that the allowed energy bands are exactly the same of a free–electron, but repeated in the k-space by reciprocal lattice vectors, $E_n(k) = \frac{\hbar^2}{2m_e}(k-nG)^2$ where $n = 0, \pm 1, \pm 2, ...$ Note the points of degeneracies: these points can be split to open gaps by a periodic potential.

8.6 Non-periodic potentials and scattering

We make a few observations and connections of the material covered in this chapter. We learn in introductory electronics courses that a signal $g_1(t) = \cos(\omega_1 t)$ and $g_2(t) = \cos(\omega_2 t)$ when mixed (or multiplied) produce signals at the sum and difference frequencies $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$ as shown in Fig. 8.14. If the initial signals were not at sharp frequencies but spread in a band, then the mixing results in sidebands that may or may not overlap.

The electron wave going through a periodic potential of a crystal is much the same problem. The Schrödinger equation $[-\frac{\hbar^2}{2m_e}\frac{\partial^2}{\partial x^2} + V(x)]\psi(x) = E\psi(x)$ for a periodic potential $V(x) = \sum_G V_G e^{iGx}$ can be

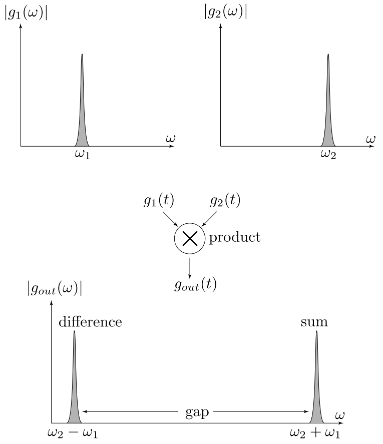


Fig. 8.14 Mixing of electronic signals produces sum and difference frequencies. The problem of electron wave propagation in a crystal is an analogous problem.

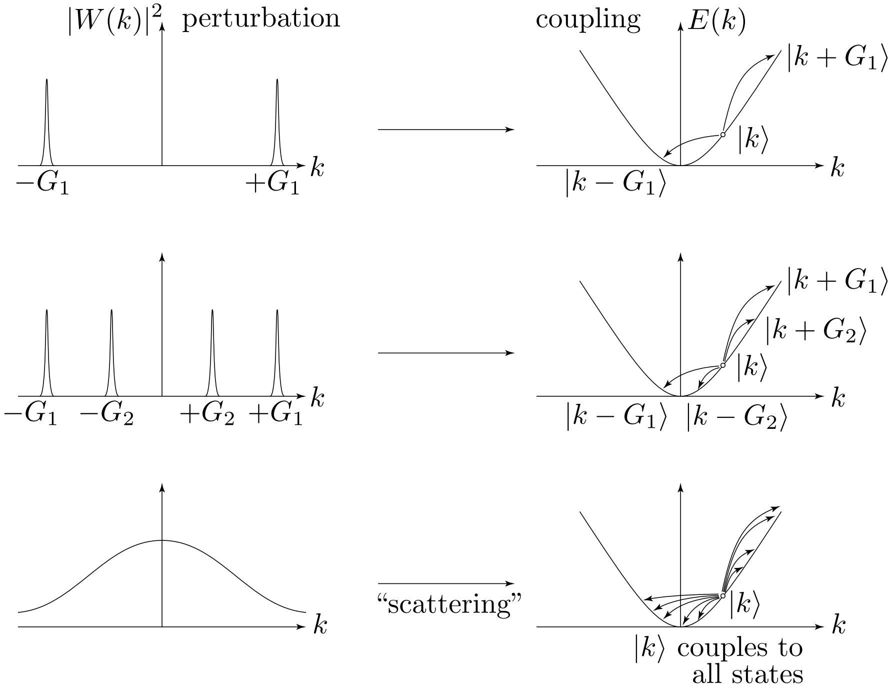

$|W(k)|^2$
perturbation
k
$-G_1$
$+G_1$
coupling
$E(k)$
$|k+G_1\rangle$
$|k\rangle$
$|k-G_1\rangle$
k
k
$-G_1$
$-G_2$
$+G_2$
$+G_1$
$|k+G_1\rangle$
$|k+G_2\rangle$
$|k\rangle$
$|k-G_1\rangle$
$|k-G_2\rangle$
k
k
“scattering”
$|k\rangle$
k
$|k\rangle$ couples to all states

Fig. 8.13 Periodic potentials only scatter states separated by specific G values, and thus open bandgaps at specific k values because they have spectral weight only for specific k's. Non-periodic potentials on the other hand can scatter a state $|k\rangle$ into several states depending on the weight of the potential in the k-space.

considered to mix the free–electron wavefunction e^{ikx} and the crystal potential to produce electron states of the form $\sum_G c_G e^{i(k\pm G)x}$.

The application of a periodic potential $-2U_G \cos(Gx)$ of reciprocal lattice vector G could only directly couple states that satisfy $k_2 - k_1 = \pm G$. This caused the appearance of bandgaps due to *direct* interaction at states $|k\rangle = |\pm G/2\rangle$. But due to *indirect* interactions, bandgaps also appeared at $|\pm mG/2\rangle$. If the periodic potential instead was $W(x) = -2[U_{G_1}\cos(G_1 x) + U_{G_2}\cos(G_2 x)]$, we expect direct gaps at more k-points, and more direct and indirect coupling of states. The nature of the periodic potential will thus determine the bandstructure. We show this schematically in Fig. 8.13.

Considering the Schrödinger equation

$$[-\frac{\hbar^2}{2m_e}\frac{\partial^2}{\partial x^2} + V(x)]\psi(x) = E\psi(x) \tag{8.31}$$

again, we note that on the left hand side, the sum of the kinetic and potential energy terms ”compete” to add up to the total energy. For wavevectors k far from $n\frac{G}{2}$, there is no strength in the crystal potential to mix electron states. Hence, the kinetic energy wins, and these electron states behave *almost* as if they are free–electrons unaware of the crystal potential. But when $k \approx \pm n\frac{G}{2}$, then the crystal potential has enough strength, the potential energy wins, electron states are not allowed. This is another view of the appearance of gaps at $k \approx n\frac{G}{2}$ that separate allowed electron states that can propagate.

If instead of a periodic potential, we had a localized potential, say $W(x) = V_0 e^{-x/x_0}$, then we can Fourier-expand the potential to obtain $W(x) = \sum_G U_G \cos(Gx)$, and the expansion coefficients will dictate the strength of the couplings. We immediately note that since a localized potential will require a large number of G's, it will effectively couple a wide range of k-states, as shown in Fig. 8.13. This is why any *deviation* from periodicity will couple a continuum of k-states, a phenomena that is responsible for **scattering** and **localization**.

Applications of non-degenerate and degenerate perturbation theory can explain a host of phenomena in semiconductors, and other quantum systems. In this chapter, we applied the techniques to the "toy-model" of an electron in a 1D periodic potential. In the next chapter, we investigate this technique with the Bloch theorem to develop useful models for the electronic bandstructure of realistic semiconductors.

8.7 Chapter summary section

In this chapter, we learned:

- A free–electron can occupy states that have a continuum of energy eigenvalues.
- In a periodic potential, free–electron states whose wavelengths are resonant with the periodic potential undergo diffraction and scattering, and form standing waves.
- The strong interaction opens bandgaps by forbidding propagating states of electrons within a band of energies, opening bandgaps.
- The total number of states do not change due to the periodic potential. Thus, the states bunch up, and the bands curve near the edges of the gap.
- Filled bands carry zero net current.
- If the Fermi level is in the middle of an allowed band, the crystal is a metal, and if it is in a gap, the crystal is an insulator.
- The curvature of the bands near the gap are called the effective masses.
- Because of level repulsion, the larger the gap, the larger the effective mass.
- If the Fermi level is near the bottom of an allowed band, it is a n-type conductor. If the Fermi level is near the top of an allowed band, it is a p-type conductor. The Hall effect signs of the two types of crystals are opposite.

- Because periodic potentials couple state $|k\rangle$ with several other states $|k \pm G\rangle$ where G is the reciprocal lattice vector, the new allowed states of the electron take the form of a Bloch function.
- Periodic potentials scatter specific electron states because of the large spectral potential at specific $k = G$.
- Non-periodic, localized, or random potentials scatter electrons into several states.

Further reading

There are a wealth of excellent books written on the band theory of solids, each covering the rich subject from different angles and perspectives. **Quantum Theory of Solids** by R. Peierls, one of the discoverers of the formation of bands and gaps is a timeless reference. A lucidly written book that should find more usage than it does is **Principles of the Theory of Solids** by J. Ziman. The same can be said of **The Theory of the Properties of Metals and Alloys** by N. Mott and H. Jones. For the historical development of the subject, a very readable account is in the article *The development of the quantum-mechanical electron theory of metals: 1928-1933* in the journal **Reviews of Modern Physics**, volume 59, page 287, 1987 by L. Hoddeson, G. Baym, and M. Eckert. **Electronic Semiconductors** by E. Spenke is an early comprehensive book that applied the concepts of condensed matter physics to semiconductors in a pedagogical manner.

For a very comprehensive discussion of the topics in this chapter, addressing the problem from various angles, I highly recommend the two references **Solid State Physics** by N. Aschroft and D. Mermin, and **Quantum Mechanics for Engineering, Materials Science, and Applied Physics** by H. Kroemer.

Exercises

(8.1) **Miscellaneous questions**
Give short answers to the following questions. All symbols have their usual meanings.

(a) Explain the origin of electron energy bands and gaps using a sketch of the $E(k)$ bandstructure of a free–electron, and of an electron in a periodic crystal.

(b) What determines if a crystal is a metal or a semiconductor? Do metals have more *net* electrons than a semiconductor?

(c) If I sum the expression for the quantum current $J = q \sum_k v_g(k) f(k)$ over all k states of the bandstructure in a *completely filled band*, what do I get?

(d) Show by using the quantum expression of current $J = q \sum_k v_g(k) f(k)$ that the current due to an *empty* electron state at the top of a valence band flows in a direction *opposite* to *filled* electron states at the bottom of the conduction band. What does this imply about the *effective charge* of a hole?

(e) Explain why adding more electrons to a n-type semiconductor will *increase* its electrical conductivity, but doing the same for a p-type semiconductor will *reduce* its electrical conductivity.

(8.2) **Electrons get their bands and gaps**

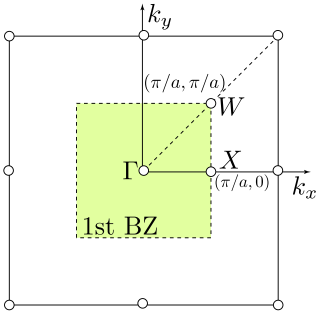


Fig. 8.15 2D reciprocal lattice.

As shown in Fig. 8.15, in the **k**-space of a 2D square lattice (lattice constant: a), denote the points $\Gamma : (k_x, k_y)$=(0,0), $X : (\pi/a, 0)$, and $W : (\pi/a, \pi/a)$. The nearly free–electron bandstructure assumes no crystal potential, but a lattice.

(a) Draw the nearly free–electron bandstructure from the BZ center in the $\Gamma - W$ direction slightly beyond the BZ edge. Identify the magnitude of k at the BZ edge, and express the energy in terms of $F = \hbar^2\pi^2/ma^2$. Include reciprocal lattice vectors smaller than $2 \times 2\pi/a$.

(b) Label each band with the reciprocal lattice vector it is associated with. Clearly point out the *degeneracies* of each band.

Consider now that the basis atoms produce a 2-D potential

$$V(x, y) = -4V_0 \cos(\frac{2\pi x}{a}) \cos(\frac{2\pi y}{a}). \qquad (8.32)$$

(c) Find the bandgap at the W point due to this potential. Be judicious in choosing the basis set.

(d) The lowest energy at the $\Gamma : (k_x, k_y) = (0, 0)$ point *before* the potential was turned on was $E_\Gamma(0, 0) = 0$ eV. Give an estimate of the change in this energy eigenvalue due to the periodic potential.

(8.3) **1D DOS, gaps, and band-edge effective masses** In this problem, we evaluate the density of states and some properties of the band-edge effective masses of electron states for the 1D crystal. The effect of a periodic potential on the energy dispersion near the energy gap for an electron was found in Equations 8.18 and Fig. 8.8, which are needed for this problem. The expressions for the band-edge effective masses $m_c^\star$ and $m_v^\star$ were obtained in Equation 8.19.

(a) Evaluate the density of states of the free–electron in the absence of the periodic potential and sketch it as a function of energy. This is identical to what was obtained in Equation 5.20 of Chapter 5.

(b) Show that the density of states only *near* the conduction band edge is given by

$$g_{1d}^c(E) = g_s \sqrt{\frac{2m_c^\star}{\hbar^2}} \frac{1}{\sqrt{E - E_c}}, \qquad (8.33)$$

and for the valence band edge by

$$g_{1d}^v(E) = g_s \sqrt{\frac{2|m_v^\star|}{\hbar^2}} \frac{1}{\sqrt{E_v - E}}. \qquad (8.34)$$

Note that these DOS are identical in expression to the free–electron DOS, but their origins are the band edges, and the masses are the effective masses. Both changes are the result of the periodic potential. This result carries over to higher dimensions.

(c) Overlay the band-edge DOS of part (b) on the free–electron DOS of part (a). Note that the band-edge DOS does not extend too far in energy from the band edges. Explain the differences due to the perturbation.

(d) Qualitatively extend the 1D DOS over the entire bands in the overlay of part (c). Explain the characteristics. If I integrate the perturbed DOS over any complete band of energies, what is the result?

(e) If $E_c - E_v = E_g = 1.0$ eV, $m_c^\star \approx 0.2m_e$, and $m_v^\star \approx -0.33m_e$, and the Fermi level E_F is exactly at the middle of the bandgap, estimate the 1D density of electrons in the conduction band, and holes in the valence band at $T = 300$ K.

(e) Show that for the 2-band problem, the inverse sum $\frac{m_e}{m_c^\star} + \frac{m_e}{m_v^\star} = 2$, i.e., it is equal to the number of bands. If I tell you that for the two-band problem I have measured the conduction band edge effective mass to be $m_c^\star \approx 0.2 m_e$, then what is $m_v^\star$? This is a special case of a more general **effective mass sum rule**: for the N-band problem, $\sum_{n=1}^{N} \frac{m_e}{m_n^\star} = N$. We will derive this relation in Chapter 12.

(8.4) **Magnitudes of bandgaps and effective masses for the electron in a 1D crystal**
In this problem, assume that the periodic potential the electron experiences in the 1D crystal is $V(x) = -2U_G \cos(Gx)$, where $U_G = 1$ eV, and $G = \frac{2\pi}{a}$ where the lattice constant (or length of a unit cell) is $a = 0.5$ nm.

(a) Make a plot of the periodic potential. Determine the classical energies allowed for an electron in this potential.

(b) Calculate the original energy of the free–electron where the bandgap opens.

(c) Calculate the energy bandwidth of the lowest electron band, and the energy bandgap that separates the lowest band from the next band.

(d) Calculate the effective masses at the top of the lowest band, and at the bottom of the second band.

(e) If the Fermi level E_F is exactly at the middle of the bandgap, estimate the 1d density of electrons in the conduction band, and holes in the valence band at $T = 300$ K.

(f) (Hard): What is the lowest energy allowed for the electron in the presence of the periodic potential?

(8.5) **Matrix calculation of a 1D bandstructure**
Using the matrix method for a large number of k-states, Calculate the energy bandstructure and density of states for electrons in a 1D periodic potential $W(x) = -2U_G \cos(Gx)$ where $U_g = 4$ eV and $G = 2\pi/a$ with a lattice constant $a = 5$ Å. Find the first and second energy gaps. If each atom has 4 electrons, find the energy bandgap and the conduction and valence band effective masses.

(8.6) **Valence bands and hole conduction**
Consider a 1D p-type conductor with the Fermi level near the top of an allowed band, where the effective mass is $m_v^\star$, and the top of the band is at $k = 0$. Answer the following questions.

(a) Sketch the bandstructure $E_v(k) \approx E_v + \frac{\hbar^2 k^2}{2m_v^\star}$. Because the band curves **downwards**, argue that m_v must be negative.

(b) From a transport measurement, you have concluded that the mobile 1D hole density in this band is p_{1D}. Show how you can find the Fermi level E_F with respect to the valence band E_v.

(c) Consider an *electron* in this band, at an energy state far below $E = E_F$. If an electric field is applied, why can it not move?

(d) Consider an *electron* in this band, at an energy state very close to $E = E_F$. If an electric field is applied, why can it move? Show why the movement of an electron in response to the electric field is equivalent to the movement of a hole in the opposite direction in the k-space.

(e) From the above parts, show why one can completely describe the transport in this band by not considering the electrons, but only holes.

(8.7) **Metal, insulator or semimetals**
Fig. 8.16 shows two bandstructures of electrons in a crystal in the reduced zone. (a) is the bandstructure of a 1D crystal, whereas (b) is that of a 2D or 3D crystal. Answer the following questions:

(a) If there are N atoms in the 1D crystal of bandstructure in Fig. 8.16 (a), and each atom has two electrons, is it a metal or a semiconductor?

(b) If by field effect I manage to put $1.8 \times N$ electrons in the 1D crystal, is it a n-type or a p-type conductor? What if I put $2.2 \times N$ electrons?

(c) Prove that in 1D, there can be no overlaps of bands as shown in Fig. 8.16 (b). What causes the overlaps in higher dimensions? (Hint: consider the nearly free–electron model of bandstructure shown in Fig. 8.12 and consider its higher-dimensional analogues.)

(d) Now the result that each band in the first Brillouin zone has exactly N states carries over to higher dimensions. If I put exactly $2 \times N$ electrons in the material that has the bandstructure of Fig. 8.16 (b), argue why the lowest band will **not** be completely filled. Sketch where you expect the Fermi level should be.

(e) For part (d), explain why this material has both electrons and holes to conduct current (in the language of semiconductor physics, this material has both electron and hole pockets in the k-space). Such materials are called **semimetals**, and are characterized by a Fermi level that is in a region of overlapping bands in regions of opposite curvatures.

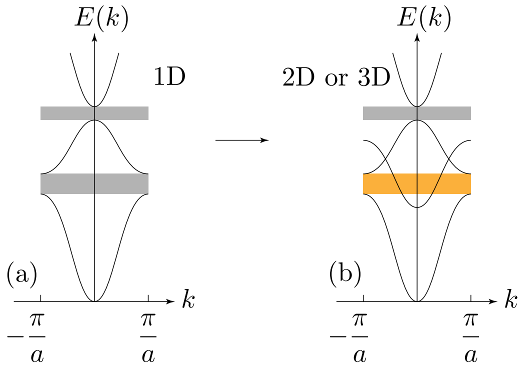


Fig. 8.16 Impossibility of band overlaps in 1D, and possible overlaps in 2 and 3D crystals.

(8.8) **Periodicity of bands in the k-space, Bloch oscillations, and tunneling**

In the discussion of the energy bandstructure in the "repeated zone" scheme in Fig. 8.8, it was argued that after the opening of the bandgap, the $E(k)$ should become periodic in k-space with periodicity $G = \frac{2\pi}{a}$ where a is the lattice constant.

(a) Using the discussion in Section 8.5 argue why $E(k) = E(k + n \cdot G)$. This means that the energy bandstructure is periodic in the k-space.

(b) Fig. 8.17 shows the energy bandstructure plotted on a cylinder, by wrapping the linear bandstructure. This way of plotting makes it clear that the band-edge states $1 = 2$ and $3 = 4$. If an electron is in the lower band, describe its motion in the curved k-space and in the real space x-axis in response to a constant force F_0 assuming the electron stays in that band.

(c) In part (b), you will find that the motion of the electron in real space (and k-space) is oscillatory. This phenomena is called **Bloch oscillations**. What is the frequency ω_B of this oscillation? Estimate its numerical value for $a = 0.5$ nm.

(d) Show that the frequency of Bloch oscillations ω_B is independent of which band the electron resides. Argue on the other hand why the real space oscillation amplitude does depend on the band the electron is in.

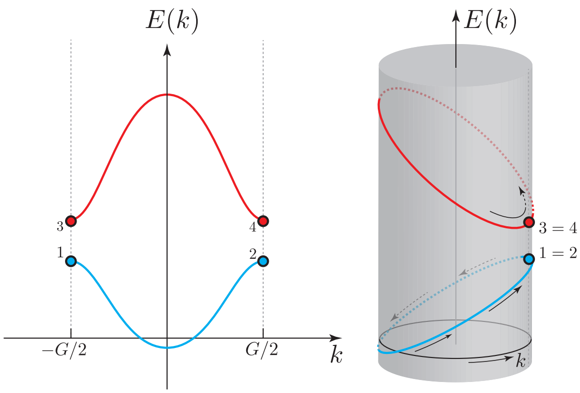


Fig. 8.17 1D bandstructure plotted in a periodic k-space.

(e) Electrons stay within respective bands for moderate strengths of electric fields. At very high fields, electrons can make a transition from the lower to the higher bands. This phenomena is called **interband tunneling** because electrons go through the forbidden energy barrier equal to the bandgap, which for the lowest bands in the 1D periodic potential $V(x) = -2U_G \cos(Gx)$ is $2U_G$. Argue why this should occur when the energy corresponding to the Bloch frequency $\hbar\omega_B$ becomes of the order of the bandgap $2U_G$. Show that the characteristic electric field for interband tunneling is then $F \approx 2U_G/a$. Estimate this field for a gap of 1 eV and a lattice constant $a = 0.5$ nm. This a *very* crude estimate, we will refine it in Chapter 24.

(8.9) **Peierls' distortion**

Fig. 8.18 shows an instability in a 1D crystal that was pointed out by Rudolph Peierls (Fig. 8.5) in his initial analysis of the energy before and after the opening of gaps. Peierls instability is a situation when a 1D crystal that should behave as a metal actually distorts and becomes an insulator. Answer the following questions, all at $T = 0$K to understand his argument.

(a) Consider the left figure. Assume the energy bandstructure is given by $E(k) = E_0[1 - \cos(ka)]$. The lattice constant is a, and there is one electron per atom. Show that $k_F = \frac{\pi}{2a}$.

(b) Find the **total energy** of all electrons in this crystal. Is it a metal or an insulator?

(c) Now consider the figure on the right. The atoms have paired (or dimerized), and the new lattice constant has become $2a$. Show why the bandstructure will be as shown, and argue why if this happens, the result is an insulator.

(d) Let us say that the total energy cost to dimerize is $\mathcal{U}_d$. Note that the dimerization energy cost is a lattice energy cost, not a direct electron energy cost. Assume the new bandstructure of the lowest band is $E_{lower}(k) = E_1[1 - \cos(2ka)]$, and the next band is separated from this band by a gap. Find E_1 so that the Peierls distortion is energetically favorable. To solve this problem, you must again find the total electronic energy with the new bandstructure.

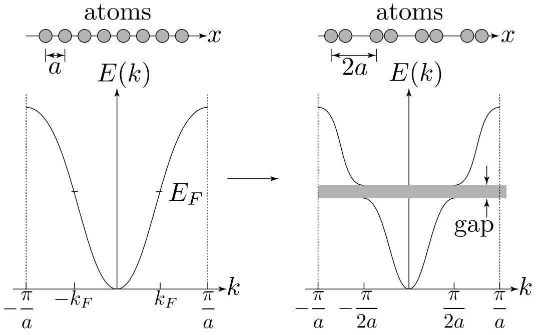


Fig. 8.18 The Peierls instability in 1D crystals.

Bloch Theorem, Bandstructure, and Quantum Currents

9

9.1 The Bloch theorem 176
9.2 Bloch theorem: aftermath 179
9.3 Real and reciprocal lattice, Brillouin zones 182
9.4 Velocity of Bloch states 187
9.5 Dynamics of Bloch states 190
9.6 Bloch wave velocity and ballistic current 194
9.7 Transport by Bloch waves with scattering 195
9.8 Energy (heat) current 197
9.9 Any current 198
9.10 Quantum Wiedemann–Franz law 199
9.11 Metals, semiconductors, semimetals and insulators 199
9.12 Chapter summary section 199
Further reading 201
Exercises 201

In Chapter 8, we saw that when an electron is placed in a periodic crystal potential, the wavefunction allowed by the Schrödinger equation is, not surprisingly, no more that of a free–electron. The philosophy of that chapter was that a crystal potential of lattice constant a and reciprocal lattice vectors $G = 2\pi/a$ *scatters* a free–electron state of wavevector k to and from states of wavevectors $k \pm G$. Because of the scattering, the free–electron wavefunction $\psi_k(x) = \frac{e^{ikx}}{\sqrt{L}}$ is perturbed to the form $\psi(x) = e^{ikx}u_k(x)$ (Section 8.5). Here $u_k(x+a) = u_k(x)$ is the periodic modulation of the free–electron wave bestowed on the electron state by the periodic lattice. The application of perturbation theory to identify the nature of the eigenfunction may give the impression that it is only approximate and probably numerically inaccurate. So what are the *exact* eigenfunctions and energies of the Schrödinger equation for the electron in the presence of a periodic potential? It turns out that we have already peeked at the prize! The Bloch wavefunction $\psi(x) = e^{ikx}u_k(x)$ happens to be the *exact* form of the eigenfunction for electron states in a periodic crystal. It is the gateway to explaining the remarkably rich range of physical phenomena born out of the marriage of **periodicity** and **quantum mechanics**.

The goals of this chapter are to understand:

- The Bloch theorem: the **exact eigenfunctions** of an electron in a periodic crystal are $\psi_k(x) = e^{ikx}u_k(x)$, where $u_k(x+a) = u_k(x)$ has the same periodicity as the crystal.
- What is the **crystal momentum** $\hbar\mathbf{k}$, and a **Brillouin zone**?
- How can Bloch theory enable the identification of **energy bandstructure** and DOS of electron states in crystals?
- What is the **velocity** and **current** carried by Bloch states? How do these respond to **external forces** and **scattering**?
- Why are metals, semiconductors, and insulators different electronic phases of matter? Why do some divalent metals have *lower* electrical conductivities than their monovalent neighbors?

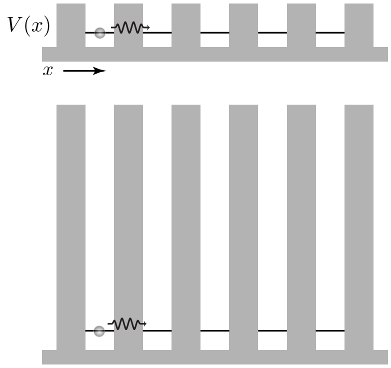


Fig. 9.1 Hopping of electrons from one atomic potential well to another for different barrier heights is not enough to explain the $\sim 10^{30}$ range of electronic conductivities seen from metals to insulators.

As was discussed in Exercise 2.1, electrical conductivities σ range from the best metals to the best insulators over a dizzying range of $\sim 10^{30}$. Fig. 9.1 shows an attempt to explain this range of conductivities *without* Bloch theorem. In an electrical insulator, the electrons could be bound strongly to the atoms, and the internal electric fields and potentials could be strong enough to keep the electrons stuck to the atoms. High electrical conductivity is explained qualitatively if the fields and barriers bind electrons only weakly to the atoms. An electron bound to an atom faces a potential barrier to hop to the next atom. Classically, it must go *over* the barrier, but quantum mechanics allows the electron to *tunnel* through the barrier. Typical interatomic distances are a few Å for both metals and insulators. If the barrier heights are small in the metal, and large in the insulator, can the tunneling probability be different by a factor of 10^{30}? In Chapter 24, we will calculate that for typical differences in barrier heights of a few eV, the tunneling probabilities can be different by perhaps a factor of $\sim 10^3$ or so. This falls *way* short of the experimentally measured range! A conceptually new idea is required. The Bloch theorem, and the concept of Brillouin zones provides that conceptual leap.

9.1 The Bloch theorem

The Bloch theorem states that the allowed eigenfunctions for electrons in a periodic potential $V(x+a) = V(x)$ of lattice constant a take the form $\boxed{\psi_k(x) = e^{ikx}u_k(x)}$, where $u_k(x+a) = u_k(x)$ shares the periodicity of the lattice, and k is *any* wavevector. We will now obtain this result quantitatively. Assume without loss of generality that the electron is on a ring of circumference L and experiences the periodic potential $V(x+a) = V(x)$. There are N lattice sites such that $L = Na$ as shown in Fig. 9.2. Now due to **symmetry**, the probability distribution $|\psi(x)|^2$ of any allowed state of the electron on the ring has no choice but to be periodic with the lattice constant, implying

$$|\psi(x+a)|^2 = |\psi(x)|^2 \implies \psi(x+a) = e^{i\phi}\psi(x)$$
$$\implies \boxed{\psi_k(x+a) = e^{ika}\psi_k(x).} \tag{9.1}$$

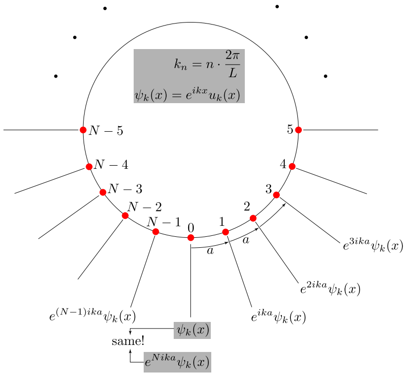


Fig. 9.2 Bloch functions

For the boxed equation, we have assumed the phase $\phi = ka$, where k is simply a variable with no physical meaning as of now. But for every chosen value of k, the wavefunctions that satisfy the Schrödinger equation become linked (or *indexed*) by k. A physical meaning for k emerges as soon as we enforce the periodic boundary condition with $L = Na$:

$$\psi_k(x+2a) = e^{i2ka}\psi_k(x), ..., \psi_k(x+Na) = e^{iNka}\psi_k(x) = \psi_k(x)$$
$$\implies e^{iNka} = 1 \implies k_n = n\frac{2\pi}{L}. \tag{9.2}$$

Fig. 9.2 shows this schematically. The boxed part of Equation 9.1 is therefore a statement of the Bloch theorem. It says that in a periodic potential, *if we translate x by a, then the Bloch wavefunction acquires a phase factor e^{ika}, where for electrons in a crystal, $k = n\frac{2\pi}{L}$*. But this is not exactly the form in which the Bloch theorem is usually stated.

To get to the usual form[1], let $\psi_k(x) = e^{ikx}u_k(x)$ be an allowed eigenfunction for an electron in a periodic potential $V(x+a) = V(x)$, where k is an allowed wavevector and $u_k(x)$ is a function split out of the free–electron wavefunction part. Then,

$$\psi_k(x+a) = e^{ik(x+a)}u_k(x+a) = e^{ika}(\underbrace{e^{ikx}u_k(x+a)}_{\psi_k(x)=e^{ikx}u_k(x)})$$
$$\implies u_k(x+a) = u_k(x), \qquad (9.3)$$

which proves that the allowed electron wavefunctions for electrons in a periodic potential $V(x+a) = V(x)$ are

$$\boxed{\psi_k(x) = e^{ikx}u_k(x),} \text{ where } \boxed{u_k(x+a) = u_k(x).} \qquad (9.4)$$

In the above expressions, $k = n\frac{2\pi}{L}$ is an electron wavevector. This form of the Bloch eigenfunction is that of a free–electron e^{ikx}, but modulated by a function $u_k(x)$ that has the same periodicity as the lattice. This is schematically shown in Fig. 9.3. It is this form that makes a direct connection to a free–electron wavefunction, which is simply e^{ikx}. If there is no crystal potential, $V(x) = 0$. Then $u_k(x) =$ constant in space, and the free–electron wavefunction is recovered. However, taking the crystal potential *gradually* to zero while always preserving the symmetry leads to $u_k(x) \to e^{iGx}$, where G is a reciprocal lattice vector – we shall see this shortly. So we can say that in the Bloch form of the wavefunction $\psi_k(x) = e^{ikx}u_k(x)$, the term e^{ikx} accounts for the crystal, or global properties, and the periodic part $u_k(x)$ accounts for the local, or atomic properties. One can consider that the periodic lattice perturbs the free–electron wavefunction (see Chapter 8, Section 8.5) and the plane wave function acquires a part $u_k(x)$ that has the same periodicity as the lattice:

$$e^{ikx} \to e^{ikx}(\underbrace{1 + \sum c_G(k)e^{iGx}}_{u_k(x)}). \qquad (9.5)$$

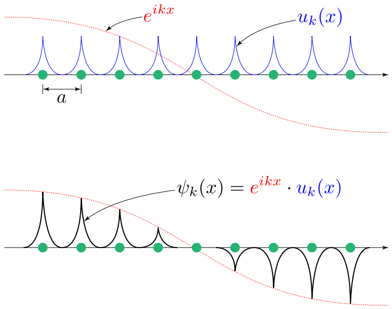


Fig. 9.3 The Bloch function is a plane wave modulated by a function periodic in the lattice constant. Note that the plots are for the real (or imaginary) parts *only*, because the probability density $|\psi_k(x)|^2$ must be periodic in the lattice constant.

This result is truly remarkable, because it explains why electrons can propagate in a crystal with a periodic potential **without scattering**. From Section 8.6, the amplitude of scattering from a state $|k\rangle \to |k'\rangle$ in the periodic potential $V(x+a) = V(x)$ of the crystal is determined by the matrix element

$$\langle k'|V(x)|k\rangle = \int_0^L dx \cdot e^{i(k'-k)x} \cdot [\underbrace{V(x)u^\star_{k'}(x)u_k(x)}_{\sum_G c_G e^{iGx}}] = \sum_G c_G \delta_{k'-k,G}, \quad (9.6)$$

[1] The general mathematical result for periodic functions due to Gaston Floquet, known as the Floquet theory was known a few decades before Bloch made the connection between quantum mechanics in spatially periodic potentials.

where $u_k(x), u_{k'}(x)$, and $V(x)$ are all periodic with the lattice constant a, and can be expanded in terms of the same reciprocal lattice vector $G = 2\pi/a$. The matrix element above is zero unless the k, k' are separated by a reciprocal lattice vector. We had discussed this topic in Chapter 8 extensively, identifying the condition for opening of bandgaps as the **Bragg condition** $|k - k'| = G$, and $|k| = |k'|$, which is met if $|k| = G/2$. We will shortly see, now from a non-perturbative (exact method), that the locus of k-points that meet this criterion define the boundaries of **Brillouin zones**. From perturbation theory, the energies of the states at these boundaries are modified the most, because the denominator of the perturbation $|V_G|^2/(E(k) - E(k'))$ vanishes. States with k near the Brillouin zone boundaries interfere strongly with the lattice, forming standing waves and energy gaps[2]. Electron states with wavevectors far away in the k-space from the Brillouin zone boundaries propagate in the crystal almost as free–electrons, except their probability densities are not constant in space, but satisfy the periodicity of the lattice because $|\psi_k(x+a)|^2 = |u_k(x+a)|^2 = |u_k(x)|^2$. We are now ready to let go of the perturbative approach, and focus on the exact Bloch picture. We immediately realize some new facts:

[2] The role of Brillouin zones in a crystal is rather similar to the role of orbitals and shells in atoms. Filled Brillouin zones will carry zero current and lead to insulators, similar to the filled outer electron shells of inert gases. Partially filled Brillouin zones will lead to electronic conductivity in metals and semiconductors, similar to the high chemical reactivity of atoms with partially filled outer shells.

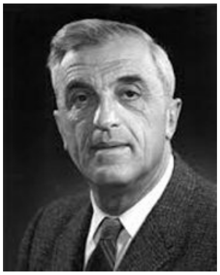


Fig. 9.4 Felix Bloch showed mathematically that electron waves can propagate in a crystal with no scattering, by introducing a wavefunction that electrons experiencing a periodic potential must satisfy. Bloch was awarded the Nobel Prize in Physics in 1952 for his work on nuclear magnetic resonance.

- Because Bloch states are *eigenstates* of the Schrödinger equation with the periodic potential, an electron in such a state is a **stationary state**: it will stay in that state forever. If the electron in a Bloch state was moving at a certain velocity, it will maintain its velocity forever in spite of the periodic lattice.
- This quantum mechanical picture turns the physics of electrical conductivity of a periodic crystal on its head. Instead of the semi-classical pictures of Drude and Sommerfeld where the finite conductivity of metals (instead of *zero* conductivity) was explained by invoking infrequent scattering of electron states from the crystal lattice, our job after this quantum revelation is to explain why the electrical conductivity of a periodic crystal is not *infinite*!

Even Bloch, and his advisor Heisenberg found this result rather remarkable.[3]

---

[3] "When I started to think about it, I felt that the main problem was to explain how the electrons could sneak by all the ions in a metal so as to avoid a mean free path of the order of atomic distances. Such a distance was much too short to explain the observed resistances... To make my life easy, I began by considering wave functions in a one-dimensional periodic potential. By straight Fourier analysis *I found to my delight that the wave differed from the plane wave of free–electrons only by a periodic modulation* $[\psi_k(x) = e^{ikx}u_k(x)]$. This was so simple that I didn't think it could be much of a discovery, but when I showed it to Heisenberg he said right away: "That's it!" Well that wasn't quite it yet, and my calculations were only completed in the summer when I wrote my thesis on *The Quantum Mechanics of Electrons in Crystal Lattices*." [F. Bloch 1976]

9.2 Bloch theorem: aftermath

In Chapter 8, we extensively discussed the *perturbative* approach to the idea of electron **energy bands**, energy **bandgaps**, the **bandstructure** $E(k)$, **group velocities**, quantum currents, **effective masses**, and electrons and **holes**. Each one of these concepts carry over to the Bloch states, so there is no need to repeat the discussion. However, the *exact* formulation of the Bloch theorem refines several of these concepts, and sets up the methods for the evaluation of the energy bandstructure and bandgaps of real semiconductor (or metal, or insulator) crystals. We discuss these new ideas that emerge from the exact Bloch formulation in this section, and pinpoint the new concepts that were not available in the perturbative approach.

Exact Bandstructure: Now consider the Schrödinger equation for a real crystal with a periodic potential $V(r)$ due to the atoms seen by an electron. Because the potential $V(r+a) = V(r)$ is periodic with a lattice constant a, it may be expanded in the reciprocal lattice vectors $G = n \cdot 2\pi/a$, with n any integer, as $V(r) = \sum_G \mathcal{V}_G e^{iGr}$. Here $\mathcal{V}_G$'s are the complex coefficients of the potential in the reciprocal (**k**-)space. Similarly, since the periodic part of the Bloch function $u_k(r)$ has the same periodicity as the lattice, it may be written as $\psi_k(r) = e^{ikr} \cdot (\sum_G u_G e^{iGr})$ where u_G's are Fourier coefficients[4]. Introducing the Fourier decompositions of the crystal potential and of the Bloch function into the Schrödinger equation transforms it from a differential equation in real-space to a linear equation in **k**-space:

$$-\frac{\hbar^2\nabla^2}{2m_e}\Big(\sum_G u_G e^{i(k+G)\cdot r}\Big) + \Big(\sum_{G'} \mathcal{V}_{G'} e^{iG'r}\Big)\Big(\sum_G u_G e^{i(k+G)\cdot r}\Big) = E(k)\sum_G u_G e^{i(k+G)\cdot r}. \tag{9.7}$$

Here we have introduced a second index G' to keep the sums most general. Applying the Laplacian operator, filtering[5] the sums, and after re-indexing G and G', this Schrödinger equation becomes:

$$\boxed{\frac{\hbar^2}{2m_e}(k+G)^2 u_G + \sum_{G'} \mathcal{V}_{G-G'} u_{G'} = E_G(k) u_G.} \tag{9.8}$$

→ **Exact Bandstructure**, Chapter 13

Note that this is *exact*; all physical details now reside in the coefficients u_G and $\mathcal{V}_G$. The physical meaning of this representation of the Schrödinger equation in the **k**-space is transparent: the electron potential energy is affected by the crystal potential. In addition to a net kinetic energy denoted by the first term, an electron picks up the crystal potential $\mathcal{V}_{G-G'}$ by interacting with other electron states G' via the crystal potential, as depicted in Figures 8.4 and 8.9 in Chapter 8. The energy eigenvalues $E_G(k)$ must now be indexed by the reciprocal lattice vector G, and a set for various k is obtained for each G. These are the allowed energy *bands*. From our discussions in Chapters 6 and 7, we identify Equation 9.8 as a linear equation, and more specifically as

[4] To be exact, the coefficients must be written as $u_{G(k)}$ and $\mathcal{V}_{G(k)}$ since they also depend on the wavevector k. This additional index, suppressed here for simplicity, is discussed later.

[5] By filtering we mean multiplying and integrating in the form $\int dx e^{-iG'x}(...)$ and using the orthogonality $\int dx e^{-i(G-G')x} = L\delta_{G,G'}$ which filters out one term from the sum over G's. Another aid to visualization is to write the linear combinations as state vectors in the bra-ket notation and projecting as was discussed in Chapters 3, 6 and 7.

a *matrix* eigenvalue problem. The solution of this exact equation will yield the eigenvalue spectrum $E(k)$ and the values u_G, which determine the corresponding eigenfunctions $\psi_k(x) = e^{ikx} \sum u_G \cdot (e^{iGx}/\sqrt{L})$. This is the subject of Chapter 13, where we will explicitly solve the exact bandstructure problem for crystals.

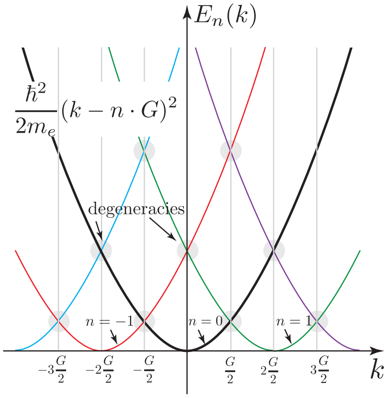


Fig. 9.5 The nearly free–electron bandstructure. The model is where there is a lattice, but no crystal potential. Bloch theorem ensures that the allowed energy bands are exactly the same of a free–electron, but repeated in the k-space by reciprocal lattice vectors, $E_n(k) = \frac{\hbar^2}{2m_e}(k - nG)^2$ where $n = 0, \pm 1, \pm 2, ...$ Note the points of degeneracies: these points can be split to open gaps by a periodic potential.

One cannot help ask what does this exact formulation tell us if we make the crystal potential vanishingly small? This is achieved by taking the limit as $\mathcal{V}_G \to 0$. Then we realize that the eigenvalues and eigenfunctions of Equation 9.8 were hiding in plain sight:

$$\boxed{E_G(k) = \frac{\hbar^2}{2m_e}(k+G)^2, \text{ and } \psi_G(x) = \frac{1}{\sqrt{L}} e^{i(k+G)x} = e^{ikx} \cdot \underbrace{(\frac{1}{\sqrt{L}} e^{iGx})}_{u(x)}.} \quad (9.9)$$

→ **Empty lattice bandstructure, Brillouin zones**, Chapter 10

Note that we precisely know the periodic part of the Bloch function here. We recover the *nearly free–electron model* bandstructure that was introduced in Chapter 8. The bandstructure is repeated again here in Fig. 9.5. Instead of a single parabola for a free–electron, the periodicity of the lattice creates several copies of it translated in the k-axis, with $k \to k + G$ where $G = n\frac{2\pi}{a}$. This is the point of departure for a precise definition of Brillouin zones, symmetry points of the crystal, and several related quantities, to which we will return in Chapter 10.

The tight-binding Bloch function: Consider a periodic array of atoms located at $x_n = na$, with $n = 1, 2, 3, ..., N$. If electrons of each atom are in an atomic orbital $\phi(x - x_n) = |n\rangle$ *before* the formation of bands, the allowed eigenfunction *after* chemical bonding is a linear combination $\psi(x) = \sum_{n=1}^{N} c_n \phi(x - x_n)$, where c_n are complex coefficients. This is guaranteed by the expansion theorem discussed in Chapter 6, because the collection of all the N atomic orbitals $|n\rangle$ forms a complete set. Now the Bloch theorem provides us a rather nice way to find the unknown coefficients c_n, and in the process provides a powerful way to find the allowed eigenfunctions $|\psi\rangle = \psi_k(x)$ and corresponding eigenvalues $E(k)$ that form the bandstructure. To see how, let us choose the coefficients $c_n = \frac{e^{ikx_n}}{\sqrt{N}}$, write out the linear combination $|\psi\rangle = \sum_{n=1}^{N} \frac{e^{ikx_n}}{\sqrt{N}} |n\rangle$, and translate $x \to x + a$ to get

$$\boxed{\psi_k(x) = \sum_{n=1}^{N} \frac{e^{ikna}}{\sqrt{N}} \phi(x - na)} \implies \psi_k(x+a) = e^{ika} \psi_k(x). \quad (9.10)$$

→ **The tight binding model**, Chapter 11.

The particular choice of the linear coefficients is thus guaranteed to be a Bloch wavefunction because it satisfies the linear translation requirement[6]. This method of solving for the electron bandstructure is called the tight-binding model. It is an intuitively satisfying, most general, and rather powerful method to calculate electronic bandstructures starting from a knowledge of atomic orbitals. Chapter 11 is dedicated to this method.

---

[6] Verify this by direct substitution $x \to x + a$ in the boxed tight-binding wavefunction, and redefining the integer indices.

The $\mathbf{k}\cdot\mathbf{p}$ method: Now consider some unique properties of the Bloch function $\psi_k(x) = e^{ikx}u_k(x)$. Notice that $\psi_k(x)$ is not periodic in x for a general k, but $u_k(x)$ is. But for the particular value $k=0$, $\psi_{k=0}(x) = u_{k=0}(x)$, meaning **the Bloch function at the center of the k-space is periodic in real space**.

Suppressing the distance x for clarity, when we apply the momentum operator on the Bloch function, we get

$$\hat{p}\psi_k = e^{ikx}(\hat{p}+\hbar k)u_k \implies (\hat{p}+\hbar k)u_k = e^{-ikx}\hat{p}\psi_k, \tag{9.11}$$

and multiple applications yield

$$\hat{p}^2\psi_k = e^{ikx}(\hat{p}+\hbar k)^2 u_k \implies ... \implies \hat{p}^n\psi_k = e^{ikx}(\hat{p}+\hbar k)^n u_k. \tag{9.12}$$

Now consider the Schrödinger equation for the periodic potential of the crystal, and generalize to any dimension:

$$[\frac{\hat{\mathbf{p}}^2}{2m_e} + V]\psi_{\mathbf{k}} = E(\mathbf{k})\psi_{\mathbf{k}}, \tag{9.13}$$

in which if we substitute the Bloch function $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r})$, and use Equation 9.12, we obtain the corresponding Schrödinger equation in the periodic function $u_{\mathbf{k}}$:

$$\boxed{[\frac{(\hat{\mathbf{p}}+\hbar\mathbf{k})^2}{2m_e} + V]u_{\mathbf{k}} = E(\mathbf{k})u_{\mathbf{k}}.} \tag{9.14}$$

This is an *effective* Schrödinger equation[7], but in the *periodic part* of the Bloch function. Note the extra $\hbar\mathbf{k}$ in the momentum operator term, which may seem like a minor change, but it has significant consequences. If we now expand the square, and use $\hat{\mathbf{p}} = -i\hbar\nabla$, we obtain the modified Schrödinger equation

$$\boxed{[-\frac{\hbar^2}{2m_e}\nabla^2 + V]u_{\mathbf{k}} + (\frac{\hbar}{m_e}\mathbf{k}\cdot\hat{\mathbf{p}})u_{\mathbf{k}} = [E(\mathbf{k}) - \frac{\hbar^2|\mathbf{k}|^2}{2m_e}]u_{\mathbf{k}}.} \tag{9.15}$$

→ The "$\mathbf{k}\cdot\mathbf{p}$" approximation, Chapter 12.

Why is this form interesting? Since $\hat{\mathbf{p}} = -i\hbar\nabla_{\mathbf{r}}$ and $\nabla^2(...) = \nabla^2_{\mathbf{r}}(...)$ are spatial derivatives, the $\mathbf{k}\cdot\hat{\mathbf{p}}$ adds to the real potential V in the Schrödinger equation. Examine the equation at $\mathbf{k}=0$: it is in a form that can be evaluated simply to yield $u_{\mathbf{k}=0}$ and $E(\mathbf{k}=0)$. This eigenvalue and eigenvector is then used to recover the related values at non-zero $\mathbf{k}$ by the use of perturbation theory of Chapter 7. This is the genesis of the $\mathbf{k}\cdot\mathbf{p}$ theory of electron bandstructure, which is very useful for semiconductors, and is the subject of Chapter 12.

---

[7] Note carefully that $\hat{H}_{\mathbf{k}} = \frac{(\hat{\mathbf{p}}+\hbar\mathbf{k})^2}{2m_e} + V$ is an effective Hamiltonian that depends on the parameter $\mathbf{k}$. This means that as we change $\mathbf{k}$, the eigenvalues $E(\mathbf{k})$ and the eigenfunction $u_{\mathbf{k}}$ change, though the equation retains the same form.

9.3 Real and reciprocal lattice, Brillouin zones

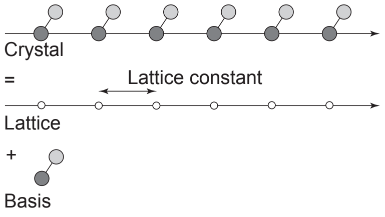


Fig. 9.6 Crystal = Lattice + Basis.

The Bloch wavefunctions of electrons in crystals are $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r})$, where $u_{\mathbf{k}}(\mathbf{r}+\mathbf{R}) = u_{\mathbf{k}}(\mathbf{r})$ is periodic with the lattice. Here, $\mathbf{R} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3}$ is a real-space **lattice vector**, and $\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}$ are the unit lattice vectors. We first precisely define what a lattice, a basis, and a crystal mean.

The **lattice** is a mathematical set of points in real space, with **no regards** to which atom or set of atoms are sitting at the lattice site. The atom, or set of atoms attached to each lattice point is called the **basis**. The mathematical symmetries present in the crystal are filtered out into the **lattice**, leaving all atomic material, chemical, and electronic details to the **basis**. We thus have the important conceptual relation

$$\boxed{\textbf{Crystal = Lattice + Basis.}} \tag{9.16}$$

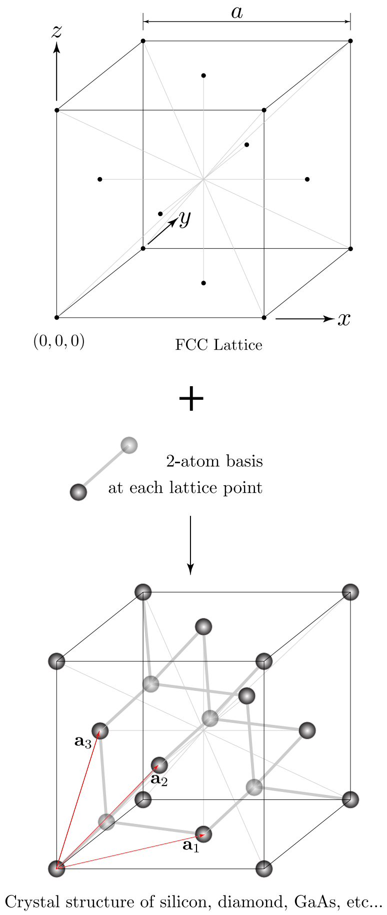


Fig. 9.7 The FCC lattice, a 2-atom basis, and the crystal of silicon, diamond, GaAs, InP, and a wide variety of semiconductors. The three arrows from $(0,0,0)$ to the face center lattice sites are the primitive lattice vectors.

The one-dimensional crystal in Fig. 9.6 highlights the above equation. The basis is the two atom dimer with the dark and light gray shades. The lattice points are uniformly spaced with a lattice constant with no regards to what atoms we decide to attach to them. The several atoms in the basis may be identical, or different elements. This means that different crystals can share the same lattice. For example, the most common semiconductor silicon has a **lattice** that is called **face-centered-cubic** (or FCC), and has a two-atom basis, both silicon. The FCC lattice, the two-atom basis, and the resulting crystal of silicon is shown in Fig. 9.7. Note carefully that though the two basis atoms are colored differently to indicate which one sits on a lattice site, they are the *same atom* in silicon. The two atoms can also be different: e.g., Ga and As, in which case Fig. 9.7 also represents a FCC GaAs crystal. Diamond, silicon, germanium, GaAs, InP, and a large number of crystals with very different electronic and photonic properties – share the same 3D lattice structure. This most common lattice structure is called ”diamond cubic” if the basis atoms are the same, and ”zinc blende” if they are different. The basis of these crystals consists of two identical atoms, such as the atoms Si and Si at coordinates $(0,0,0)$ and $a(\frac{1}{4},\frac{1}{4},\frac{1}{4})$. Similarly, 2D crystals such as graphene, BN, and transition metal dichalcogenides such as MoS_2 and NbSe_2 share a common lattice-type, but have vastly different properties!

The location of *every* lattice point of a 3D crystal is given by the vector $\mathbf{R} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3}$. This vector $\mathbf{R}$ *does not* point to any location that is *not* a lattice point. This means we can reach every lattice point of a crystal of any size – be it a few nanometers across, or a few cm across, with suitable choices of the integer triplet (n_1, n_2, n_3). The real-space **primitive lattice vectors**, $\mathbf{a_1}, \mathbf{a_2}$, and $\mathbf{a_3}$ are chosen to be the *shortest* three vectors for 3D lattices (with precise generalization to 2D and 1D crystals). Fig. 9.7 shows that the three primitive lattice vectors for silicon-like crystals are $\mathbf{a}_1 = a(\frac{1}{2},\frac{1}{2},0)$, $\mathbf{a}_2 = a(0,\frac{1}{2},\frac{1}{2})$, and

$\mathbf{a}_3 = a(\frac{1}{2}, 0, \frac{1}{2})$. There is a good reason for choosing the shortest lattice vectors. The volume spanned by the three primitive lattice vectors is the scalar triple product $\Omega_u = \mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3$. For example[8], this volume of all silicon-like crystals is

$$\begin{aligned}\Omega_u = \mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3 &= a(\frac{1}{2}, \frac{1}{2}, 0) \cdot a(0, \frac{1}{2}, \frac{1}{2}) \times a(\frac{1}{2}, 0, \frac{1}{2}) \\ &= a(\frac{1}{2}, \frac{1}{2}, 0) \cdot a^2(\frac{1}{4}, \frac{1}{4}, -\frac{1}{4}) = a^3(\frac{1}{8} + \frac{1}{8}) = \frac{a^3}{4}. \qquad (9.17)\end{aligned}$$

This volume[9] is found to be 1/4th of the volume a^3 of the conventional cell shown in Fig. 9.7. This highlights an important point: the conventional cubic cell is *not necessarily* the smallest repeating unit. The space occupied by this volume Ω_u is referred to as the **Wigner–Seitz cell**. The Wigner–Seitz cell is the *smallest* repeating unit of the crystal in *real space*. Translating the Wigner–Seitz cell by the lattice vectors $\mathbf{R} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$ must fill the entirety of the 3D space. No small nook or corner is left unfilled. The total volume of a crystal V is an integral multiple of the volume of the Wigner–Seitz cell $V = N\Omega_u$. By definition, **only one lattice point belongs to a Wigner–Seitz cell**. All information about the crystal must be present in the Wigner–Seitz cell, because every other cell is merely a repeat! As an example, the Bloch wavefunction of electrons must follow

$$\boxed{\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{R}}\psi_{\mathbf{k}}(\mathbf{r}) \implies |\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{R})|^2 = |\psi_{\mathbf{k}}(\mathbf{r})|^2} \implies$$

The Bloch wavefunction *squared* repeats in every Wigner–Seitz cell.

Since the square of the electron wavefunction represents the probability density, the **electron concentration** $n(\mathbf{r}) \propto |\psi_{\mathbf{k}}(\mathbf{r})|^2$ must have exactly the same periodicity as the lattice. As a result, if a physical property depends on $|\psi_{\mathbf{k}}(\mathbf{r})|^2$, we only have to evaluate it over one unit cell, and not the whole crystal. Now the greatest importance of the Wigner–Seitz cell is recognized when we move from the real space to the wavevector, or **k**-space. Corresponding to the three real space primitive lattice vectors are the three **primitive reciprocal lattice vectors** given by

$$\boxed{\mathbf{b}_1 = 2\pi\frac{\mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}, \mathbf{b}_2 = 2\pi\frac{\mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}, \mathbf{b}_3 = 2\pi\frac{\mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}.} \qquad (9.18)$$

With an integer triplet (m_1, m_2, m_3), the vectors

$$\boxed{\mathbf{G} = m_1\mathbf{b}_1 + m_2\mathbf{b}_2 + m_3\mathbf{b}_3} \qquad (9.19)$$

define the **reciprocal space lattice**, just as the real-space lattice of the crystal. In a 1D crystal of lattice constant a, the reciprocal lattice vector is simply $G = \frac{2\pi}{a}$. In 2D, the out of plane unit cell vector $\mathbf{a}_3$ may be chosen as $\mathbf{a}_3 = \mathbf{a}_1 \times \mathbf{a}_2$ to apply Equation 9.18. Unlike the real-space lattice where there is a basis of atoms attached to each lattice

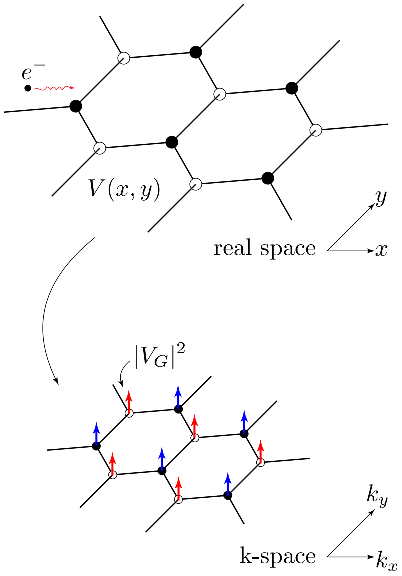


Fig. 9.8 The crystal potential in the **k**-space is **concentrated** at the reciprocal lattice points **G**.

---

[8] The crystal structures and lattice constants of semiconductors, metals, and insulators are today imaged and measured to a very high degree of accuracy using X-Ray diffraction, and Transmission-Electron Microscopy.

[9] You can use the vector identity $\mathbf{v}_1 \cdot \mathbf{v}_2 \times \mathbf{v}_3 = \mathbf{v}_2(\mathbf{v}_1 \cdot \mathbf{v}_3) - \mathbf{v}_3(\mathbf{v}_1 \cdot \mathbf{v}_2)$ for calculating triple products.

site, the reciprocal lattice points have no such attachments (Fig. 9.8). However, since the real space potential seen by electrons is periodic, it may be expanded in Fourier series as $V(\mathbf{r}) = \sum_{\mathbf{G}} \mathcal{V}_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}}$, where **G** are precisely the reciprocal lattice vectors defined by Equation 9.19. The potential $V(\mathbf{r})$ is periodic and continuous in real space, but in the **k**-space it is concentrated only at the reciprocal lattice points **G** with "strength" $|\mathcal{V}_{\mathbf{G}}|^2$, where $\mathcal{V}_{\mathbf{G}}$ is the Fourier coefficient. This is illustrated in Fig. 9.8. One can imagine the crystal potential in the **k**-space to be a 3D 'bed of nails' so to speak. Since $\mathbf{R} \cdot \mathbf{G} = 2\pi(n_1 m_1 + n_2 m_2 + n_3 m_3)$ is always an integer multiple of 2π, $e^{i\mathbf{R}\cdot\mathbf{G}} = 1$ for all real space lattice points **R** and reciprocal space lattice points **G**. The definition in Equation 9.18 implies the *lattice points in reciprocal space* are a Fourier transform of the *lattice points in real space*. Thus a square real lattice will transform into a square reciprocal lattice as shown in Fig. 9.9 (a). A rectangular real lattice thin in x and fat in y will transform into a rectangular reciprocal lattice fat in k_x but thin in k_y, and we will shortly see that a face-centered cubic (FCC) real lattice will transform into a body-centered cubic (BCC) reciprocal lattice.

Consider a periodic boundary condition of a large 1D crystal of side length $L = Na$. Similar to the arguments for the free–electron in Chapter 5, Section 5.3, we realize that the Bloch function for electrons in this crystal should satisfy the periodic boundary condition $\psi_k(x + L) = \psi_k(x)$. But the Bloch function itself has the property $\psi_k(x + L) = e^{ikL}\psi_k(x)$, implying that $e^{ikL} = 1$, and thus the allowed values of k are $k_n = \frac{2\pi}{L}n$, where $n = ... -1, 0, 1, ...$ is an integer. Note that while for the free–electron $\hbar k$ was the full momentum, for the Bloch electron, $\hbar k$ is the *crystal momentum*. This implies just as for the free–electron, the **k**-space for the electron Bloch states of a crystal is a dense grid of points separated by $\frac{2\pi}{L}$. If there are N primitive unit cells in the real space, then there are exactly $\frac{2\pi/a}{2\pi/L} = N$ discrete k-points between any two reciprocal lattice points. Similarly, the 2D **k**-points are seen as a dense grid in Fig. 9.9 (b), inside the larger reciprocal lattice grid. Inside every unit cell in the reciprocal lattice of side $2\pi/a$, there are exactly $\frac{2\pi/a}{2\pi/L_x} \cdot \frac{2\pi/a}{2\pi/L_y} = N_x N_y$ points, which again is equal to the number of unit cells in the 2D real space. In a 3D crystal, there are $\frac{2\pi/a}{2\pi/L_x} \cdot \frac{2\pi/a}{2\pi/L_y} \cdot \frac{2\pi/a}{2\pi/L_z} = N_x N_y N_z$ **k**-points in the primitive unit cell in the reciprocal space.

The effective Wigner–Seitz cell, or the smallest repeating unit in the **k**-space reciprocal lattice grid is called the **Brillouin zone** or **BZ** of a crystal. The volume of the BZ therefore is $\mathbf{b_1} \cdot \mathbf{b_2} \times \mathbf{b_3} = \frac{(2\pi)^3}{\Omega_u}$, or in other words the products of the Wigner–Seitz cells in real and **k**-spaces is $(2\pi)^3$. Similar to the free–electron model, **the number of discrete k-points in the BZ is equal to the number of unit cells in the real space**. In other words, inside the first BZ, there is one **k**-point corresponding to each unit cell in the real space, irrespective of how many atoms are in the basis, or how many electrons are attached to those atoms. **Brillouin zones are thus a lattice property, and are in-**

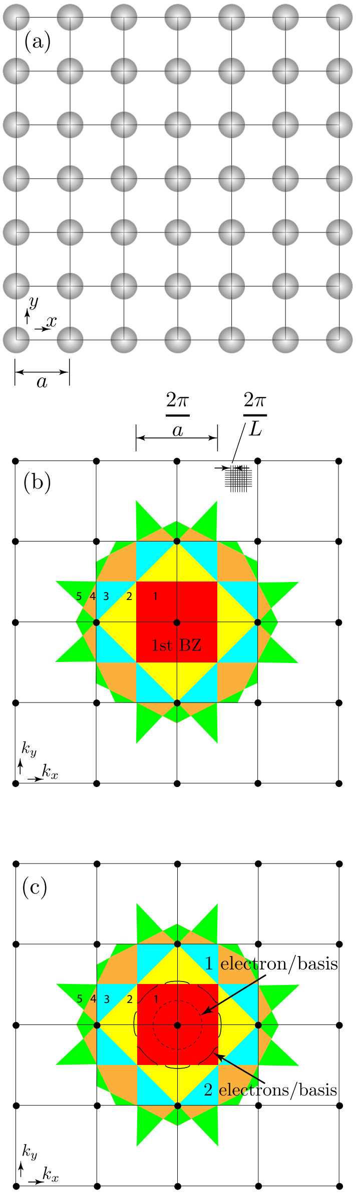


Fig. 9.9 A 2D crystal (a) and its reciprocal lattice points (b), defining the **Brillouin zones** of a 2D square lattice. Part (c) indicates that the first BZ is filled up to the dashed line constant-energy contour for monovalent atoms with $N = 1$, and spills over to the second BZ shown as solid line constant energy contours for divalent atoms with $N = 2$.

dependent of the basis. We found in Chapter 8 that electrons with $k = \pm G/2$ suffer Bragg diffraction from the crystal potential, with the resulting splitting of the energy eigenvalues, leading to the formation of energy bands separated by energy gaps. In 2D and 3D, the $\mathbf{k}$ points that undergo Bragg diffraction are located at the **edges** of the Brillouin zones. Choosing the $\mathbf{G} = (0, 0)$ $\mathbf{k}$-point as the origin, the Brillouin zones are determined by the Voronoi procedure[10]: draw the perpendicular bisector 'planes' to the nearest neighbor reciprocal lattice points (not just the first!); the smallest 'volume' enclosed by the bisectors forms the first BZ. Repeat the procedure with the next nearest neighbor reciprocal lattice points for the second BZ, and so on. Fig. 9.9 (b) indicates the first 5 BZs for the simple 2D square lattice.

Though the BZs are a lattice property, how they are filled with electrons is a property of the basis atoms and the number of electrons contributed by the basis atoms. Each $\mathbf{k}$-point can hold exactly two electrons, of opposite spins, by the Pauli exclusion principle. Thus, **each k-point in the BZ is analogous to an orbital of an atom**. If there are N atoms in the crystal, the first BZ has N $\mathbf{k}$-points, and can hold $2N$ electrons. If each atom contributes one electron to the extended Bloch eigenstates, then the first BZ is not completely filled, as indicated by the dashed circle in Fig. 9.9 (c). The energy dispersion $E(\mathbf{k})$ will determine the exact shape of the constant energy contour of the highest filled states; for parabolic bands, the contour is a circle. If the basis set of atoms contributes 2 electrons each, there are $2N$ electrons, and the filled states spill over to the second BZ, indicated by the solid contour in Fig. 9.9. The **constant energy contours are perpendicular to the edges of the BZs**, as the Bragg diffraction condition ensures that $\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}) = 0$. It is then clear that in 1D, for one electron/basis, we have a partially filled band – a metal, for two electrons each we have a filled band with a gap – an insulator (or semiconductor), for three electrons/basis we have a metal, and this behavior cycles periodically, much the same way the chemical reactivity of atoms cycles between highly reactive to inert gases, and for rather similar reasons. This simple procedure of filling of states and distinction between metals, semiconductors, and insulators is also valid for 2D and 3D, but the energy bands may overlap based on the basis atoms and crystal potentials – and therefore must be considered on a case-by-case basis, which we do in the next section.

Properties of Bloch functions and Eigenvalues in Brillouin zones: Because the grid points inside the BZ are given by $\mathbf{k} = \frac{n_1}{N_1}\mathbf{b_1} + \frac{n_2}{N_2}\mathbf{b_2} + \frac{n_3}{N_3}\mathbf{b_3}$, where $0 \leq n_1 \leq N_1 - 1$ and so on, it should be evident that if we take a sum of the nature $\sum_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{R}}$, then if $\mathbf{R} = 0$, the sum is $N = N_1 N_2 N_3$, and if $\mathbf{R} \neq 0$, the sum is zero. This is because the sum is over $N = N_1 N_2 N_3$ complex exponentials such that the sum of all the phase angles is exactly an integer times 2π, or in other words we are summing complex unit vectors around the unit circles[11]. Using the exact same argument, we also get $\sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} = 0$ if $\mathbf{k} \neq \mathbf{G}$, and $\sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} = N$ if $\mathbf{k} = \mathbf{G}$ is a reciprocal lattice vector[12].

---

[10] This mathematical procedure to demarcate areas or volumes closest to determined points found by Georgy Voronoi finds wide usage in many areas of science and engineering. Deciding which cell tower your phone should connect to is a great example.

[11] This is the same as summing the N^{th} roots of unity as discussed in Exercise 9.1.

[12] These two results are summarized as: $\sum_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{R}} = N\delta_{\mathbf{k},0}$, and $\sum_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}} = N\delta_{\mathbf{k},\mathbf{G}}$.

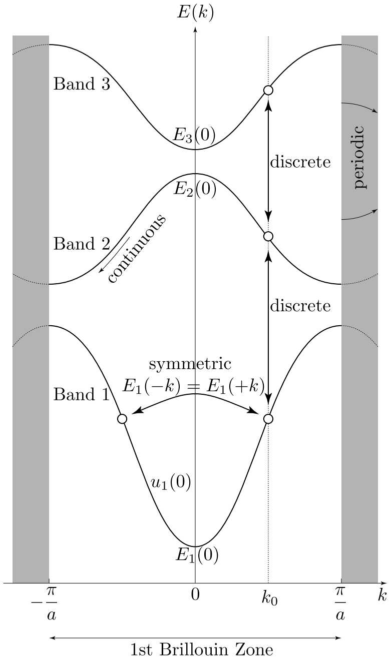


Fig. 9.10 Properties energy bands must satisfy for Bloch states in crystals.

Translating the Bloch wavefunction $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r})$ by a real space lattice vector **a** made the wavefunction acquire a phase $\psi_{\mathbf{k}}(\mathbf{r}+\mathbf{a}) = e^{i\mathbf{k}\cdot\mathbf{a}}\psi_{\mathbf{k}}(\mathbf{r})$. If we translate the wavevector **k** by a reciprocal lattice vector $\mathbf{k} \to \mathbf{k} + \mathbf{G_0}$, the Bloch function becomes

$$\psi_{\mathbf{k}+\mathbf{G_0}}(\mathbf{r}) = e^{i(\mathbf{k}+\mathbf{G_0})\cdot\mathbf{r}}\sum_{G} c_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}} = e^{i\mathbf{G_0}\cdot\mathbf{r}}(e^{i\mathbf{k}\cdot\mathbf{r}}\sum_{G} c_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}}) = e^{i\mathbf{G_0}\cdot\mathbf{r}}\psi_{\mathbf{k}}(\mathbf{r}), \tag{9.20}$$

meaning the Bloch function acquires a phase shift $e^{i\mathbf{G_0}\cdot\mathbf{r}}$. This proves that analogous to the periodicity of the squared Bloch function in real space,

$$\boxed{|\psi_{\mathbf{k}+\mathbf{G}}(\mathbf{r})|^2 = |\psi_{\mathbf{k}}(\mathbf{r})|^2.} \tag{9.21}$$

$\implies$ **The Bloch wavefunction *squared* repeats in every BZ.**

Consider the exact $\mathbf{k}\cdot\mathbf{p}$ form of the Schordinger equation for Bloch states discussed in Equation 9.16 and reproduced here:

$$[-\frac{\hbar^2}{2m_e}\nabla^2 + V]u_{\mathbf{k}} + (\frac{\hbar}{m_e}\mathbf{k}\cdot\hat{\mathbf{p}})u_{\mathbf{k}} = [E(\mathbf{k}) - \frac{\hbar^2|\mathbf{k}|^2}{2m_e}]u_{\mathbf{k}}. \tag{9.22}$$

In Fig. 9.10, consider the case for $\mathbf{k} = 0$, which is the origin in the BZ. At this **k**- point, the eigenvalues and eigenfunctions are the solutions to the much simpler Schrödinger equation

$$[-\frac{\hbar^2}{2m_e}\nabla^2 + V(\mathbf{r})]u_0(\mathbf{r}) = E(0)u_0(\mathbf{r}). \tag{9.23}$$

The energy eigenfunctions $u_0(\mathbf{r})$, in addition to being **periodic**, inherit the **symmetries of the crystal potential** $V(\mathbf{r})$. The choice of any **k** requires the wavefunction to have a definite periodicity, and any $\mathbf{k} \pm \mathbf{G}$ have smaller wavelengths, but share the same periodicity. But importantly, these states are **discrete**, meaning for each **k** there are a discrete number of allowed eigenvalues $E_n(\mathbf{k})$, with $n = 1, 2, ...$ Consider the eigenvalues $E(\mathbf{k} = 0)$, they form a **discrete** set, just like the particle in a box, the harmonic oscillator, or the hydrogen atom. Call this set $E_1(0), E_2(0), E_3(0), ...$ from the lowest to higher energies. Now since **k** is a *continuous* parameter in the Hamiltonian in Equation 9.22, the eigenvalues $E_n(\mathbf{k})$ vary continuously as **k** is varied from $\mathbf{k} = 0$ to non-zero values, all the way to the BZ edges. The discreteness of $E_n(0)$, and the continuity in **k** thus forms a band $E_n(\mathbf{k})$. Consider the point $\mathbf{k} = \frac{\mathbf{G}}{2}$. Due to the Bragg condition, we have seen in Chapter 8, Fig. 8.8 that this state is actually *the same state* as $\mathbf{k} = -\frac{\mathbf{G}}{2}$! Therefore, we have $E_n(\mathbf{k} = \mathbf{G}/2) = E_n(\mathbf{k} = -\mathbf{G}/2)$. Finally, note that the complex conjugate of the Schrödinger equation $\hat{H}e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r}) = E_n(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{r}}u_{n\mathbf{k}}(\mathbf{r})$ is

$$\hat{H}e^{-i\mathbf{k}\cdot\mathbf{r}}u^{\star}_{n\mathbf{k}}(\mathbf{r}) = E_n(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{r}}u^{\star}_{n\mathbf{k}}(\mathbf{r}) \implies E_n(\mathbf{k}) = E_n(-\mathbf{k}), \tag{9.24}$$

which implies that the energy bands $E_n(\mathbf{k})$ are symmetric around $\mathbf{k} = 0$. The complex conjugation leads to a wavefunction $e^{-i\mathbf{k}\cdot\mathbf{r}}u^{\star}_{n\mathbf{k}}(\mathbf{r})$,

Table 9.1 Symmetries of Bloch eigenvalues. A general result of quantum mechanics is the Kramers degeneracy theorem, which states that for a system of half-integer total spin which has a time-reversal symmetric Hamiltonian, every energy level is at least doubly degenerate.

Time Reversal ($t \to -t$)
$E_{n,\uparrow}(\mathbf{k}) \to E_{n,\downarrow}(-\mathbf{k})$
$E_{n,\downarrow}(\mathbf{k}) \to E_{n,\uparrow}(-\mathbf{k})$

Inversion ($\mathbf{r} \to -\mathbf{r}$)
$E_{n,\uparrow}(\mathbf{k}) \to E_{n,\uparrow}(-\mathbf{k})$
$E_{n,\downarrow}(\mathbf{k}) \to E_{n,\downarrow}(-\mathbf{k})$

TR & Inversion
$E_n(\mathbf{k}) = E_n(-\mathbf{k})$

Lattice periodicity
$E_n(\mathbf{k}+\mathbf{G}) = E_n(\mathbf{k})$

which is also an allowed Bloch function with index $-\mathbf{k}$ and periodic part $u^{\star}_{n\mathbf{k}}(\mathbf{r})$, therefore an energy eigenvalue at $E_n(\mathbf{k})$ guarantees one at $E_n(-\mathbf{k})$. Actually, this is true as long as there is spatial inversion symmetry in the crystal potential, and spin is neglected. When spin is considered, the complete range of symmetries are indicated in Table 9.1. Because of the lattice periodicity, the energy eigenvalues of each Bloch band repeats in all BZs:

$$\boxed{E_n(\mathbf{k}+\mathbf{G}) = E_n(\mathbf{k}).} \tag{9.25}$$

Note that the periodicity of $E(\mathbf{k})$ in the $\mathbf{k}$-space means that we can write $E(\mathbf{k}) = \sum_{\mathbf{R}} \mathcal{E}_{\mathbf{R}} e^{i\mathbf{k}\cdot\mathbf{R}}$, where $\mathbf{R}$ is a *real* space lattice vector and $\mathcal{E}_{\mathbf{R}}$ and expansion coefficient. In fact, this form of the energy bandstructure of Bloch states is useful for the effective mass approximation of Chapter 14 and the tight-binding bandstructure of Chapter 11.

9.4 Velocity of Bloch states

In this section, we dig deeper into the nature of Bloch states and find a few features about their velocities, and their matrix elements that are useful to understand **electron-photon**, **electron-phonon**, and **electron-defect** interactions. We then examine the dynamics of Bloch states in response to **external forces** on electrons due to electric or magnetic fields. Though this section may *appear* involved, please read it carefully because the physical contents under the notational garb are both intuitive and important.

The velocity of a Bloch eigenstate of crystal momentum $\mathbf{k}$, energy eigenfunction $\psi_{\mathbf{k}}(\mathbf{r})$ and energy eigenvalue $E(\mathbf{k})$ is the expectation value of the momentum operator:

$$\mathbf{v}(\mathbf{k}) = \langle \psi_{\mathbf{k}} | \frac{\hat{\mathbf{p}}}{m_e} | \psi_{\mathbf{k}} \rangle. \tag{9.26}$$

The Bloch eigenstates $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r})$ are energy eigenstates, but are they also[13] momentum eigenstates? They are *not*, because

$$\hat{\mathbf{p}}\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}(\hat{\mathbf{p}} + \hbar\mathbf{k})u_{\mathbf{k}}(\mathbf{r}) \neq (...)\psi_{\mathbf{k}}(\mathbf{r}). \tag{9.27}$$

This is intuitively correct because the crystal potential mixes various $\mathbf{k} \pm \mathbf{G}$ plane waves. From this relation, we note that $(\hat{\mathbf{p}} + \hbar\mathbf{k})u_{\mathbf{k}}(\mathbf{r}) = e^{-i\mathbf{k}\cdot\mathbf{r}}\hat{\mathbf{p}}\psi_{\mathbf{k}}(\mathbf{r})$, which we may write in the vector notation $(\hat{\mathbf{p}} + \hbar\mathbf{k})|u_{\mathbf{k}}\rangle = e^{-i\mathbf{k}\cdot\mathbf{r}}\hat{\mathbf{p}}|\psi_{\mathbf{k}}\rangle$. Since the periodic part of the Bloch function satisfies

$$\underbrace{[\frac{(\hat{\mathbf{p}} + \hbar\mathbf{k})^2}{2m_e} + V(\mathbf{r})]}_{\hat{H}_{\mathbf{k}}}|u_{\mathbf{k}}\rangle = E(\mathbf{k})|u_{\mathbf{k}}\rangle, \tag{9.28}$$

we take the derivative (or gradient $\nabla_{\mathbf{k}}$) in the $\mathbf{k}$-space to get[14]

[13] free–electron states $\psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}$ are simultaneously energy and momentum eigenstates.

[14] The 1D version has simpler notation, verify the steps! Note the following: a) the $\hat{H}_{\mathbf{k}}$ used here is the effective Hamiltonian of the periodic part of the Bloch function $u_{\mathbf{k}}$ and not the Bloch function $\psi_{\mathbf{k}}$ itself, b) that $\nabla_{\mathbf{k}} V(\mathbf{r}) = 0$, and like ordinary differentiation, the gradient also uses the chain-rule. Closing the bracket on the left with $\langle u_{\mathbf{k}}|$ amounts to multiplying with $u^{\star}_{\mathbf{k}}$ on the left, and integrating $\int d\mathbf{r}(...)$ over all real space (not $\mathbf{k}$-space!).

$$\frac{\hbar}{m_e}\underbrace{(\hat{\mathbf{p}}+\hbar\mathbf{k})|u_{\mathbf{k}}\rangle}_{=e^{-i\mathbf{k}\cdot\mathbf{r}}\hat{\mathbf{p}}|\psi_{\mathbf{k}}\rangle}+\hat{H}_{\mathbf{k}}\nabla_{\mathbf{k}}|u_{\mathbf{k}}\rangle=\nabla_{\mathbf{k}}E(\mathbf{k})|u_{\mathbf{k}}\rangle+E(\mathbf{k})\nabla_{\mathbf{k}}|u_{\mathbf{k}}\rangle. \quad (9.29)$$

Now if we close the bracket on the left with $\langle u_{\mathbf{k}}|$, we obtain

$$\frac{\hbar}{m_e}\underbrace{\langle u_{\mathbf{k}}|e^{-i\mathbf{k}\cdot\mathbf{r}}}_{\langle\psi_{\mathbf{k}}|}\hat{\mathbf{p}}|\psi_{\mathbf{k}}\rangle+\underbrace{\langle u_{\mathbf{k}}|\hat{H}_{\mathbf{k}}}_{E(\mathbf{k})\langle u_{\mathbf{k}}|}\nabla_{\mathbf{k}}|u_{\mathbf{k}}\rangle=\nabla_{\mathbf{k}}E(\mathbf{k})\underbrace{\langle u_{\mathbf{k}}|u_{\mathbf{k}}\rangle}_{1}+E(\mathbf{k})\langle u_{\mathbf{k}}|\nabla_{\mathbf{k}}|u_{\mathbf{k}}\rangle. \quad (9.30)$$

The last terms cancel on both sides, giving us the velocity of the Bloch state[15]:

$$\boxed{\langle\psi_{\mathbf{k}}|\frac{\hat{\mathbf{p}}}{m_e}|\psi_{\mathbf{k}}\rangle=\mathbf{v}(\mathbf{k})=\frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k}).} \quad (9.31)$$

Equation 9.31 is crucial to understand much of the electronic phenomena of solid state physics. We have now quantitatively derived what was qualitatively discussed in Chapter 8: that **the group velocity of a state k is given by the slope of the $E(\mathbf{k})$ curve**. Therefore, the velocity of the electron is zero *both* at the bottom of an energy band $E(\mathbf{k})$, *and at the top of the band*. Occupied states near the bottom of the band conduct as electrons and unoccupied near the top conduct as **holes**.

Now let us consider transitions (Fig. 9.11) between different bands $E_n(\mathbf{k})$, where a band is indexed by n with eigenfunctions $|\psi_{n\mathbf{k}}\rangle=e^{i\mathbf{k}\cdot\mathbf{r}}|u_{n\mathbf{k}}\rangle$. We again use Equation 9.28 to write

$$\underbrace{[\frac{(\hat{\mathbf{p}}+\hbar\mathbf{k})^2}{2m_e}+V(\mathbf{r})]}_{\hat{H}_{\mathbf{k}}}|u_{n\mathbf{k}}\rangle=E(\mathbf{k})|u_{n\mathbf{k}}\rangle, \quad (9.32)$$

and by taking its **k**-derivative, get

$$\frac{\hbar}{m_e}\underbrace{(\hat{\mathbf{p}}+\hbar\mathbf{k})|u_{n\mathbf{k}}\rangle}_{=e^{-i\mathbf{k}\cdot\mathbf{r}}\hat{\mathbf{p}}|\psi_{n\mathbf{k}}\rangle}+\hat{H}_{\mathbf{k}}\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle=[\nabla_{\mathbf{k}}E_n(\mathbf{k})]|u_{n\mathbf{k}}\rangle+E_n(\mathbf{k})\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle. \quad (9.33)$$

Now if we close the bracket on the left with $\langle u_{m\mathbf{k}}|$ of *another* band ensured by $m\neq n$, we obtain

$$\frac{\hbar}{m_e}\underbrace{\langle u_{m\mathbf{k}}|e^{-i\mathbf{k}\cdot\mathbf{r}}}_{\langle\psi_{m\mathbf{k}}|}\hat{\mathbf{p}}|\psi_{n\mathbf{k}}\rangle+\underbrace{\langle u_{m\mathbf{k}}|\hat{H}_{\mathbf{k}}}_{E_m(\mathbf{k})\langle u_{m\mathbf{k}}|}\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle=$$
$$\nabla_{\mathbf{k}}E_n(\mathbf{k})\underbrace{\langle u_{m\mathbf{k}}|u_{n\mathbf{k}}\rangle}_{\delta_{n,m}}+E_n(\mathbf{k})\langle u_{m\mathbf{k}}|\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle. \quad (9.34)$$

Since $m\neq n$, we obtain an useful relation for the "cross-velocity", or the interband momentum matrix element

$$\boxed{\langle\psi_{m\mathbf{k}}|\frac{\hat{\mathbf{p}}}{m_e}|\psi_{n\mathbf{k}}\rangle=[\frac{E_n(\mathbf{k})-E_m(\mathbf{k})}{\hbar}]\langle u_{m\mathbf{k}}|\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle.} \quad (9.35)$$

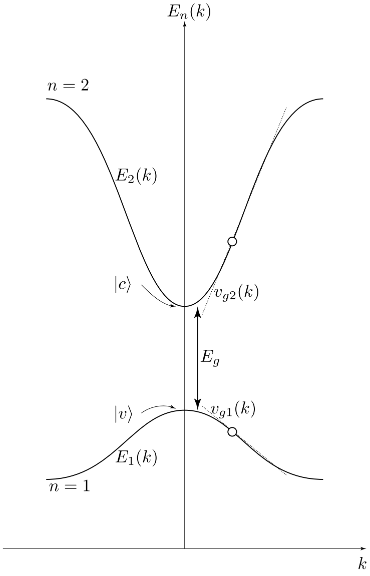


Fig. 9.11 The intra-band (or self-) momentum matrix element $\langle\psi_{n\mathbf{k}}|\hat{\mathbf{p}}|\psi_{n\mathbf{k}}\rangle$ gives the velocity of a Bloch state at any **k**. The cross-velocity, or the interband momentum matrix element $\langle\psi_{m\mathbf{k}}|\hat{\mathbf{p}}|\psi_{n\mathbf{k}}\rangle$ belonging to two bands with different reciprocal lattice vectors will be useful later for interband optical transitions in which a photon is absorbed or emitted.

---

[15] This remarkably simple and endlessly useful expression for the velocity of Bloch states can be derived from many different perspectives, see Exercise 9.3 for further insights. You may wonder: what if the shape of the bandstructure $E(\mathbf{k})$ is such that the derivative $\nabla_{\mathbf{k}}E(\mathbf{k})$ is not well defined? This is an interesting question which leads to the **Berry curvature** of Bloch states, via a modification of this constitutive equation. This is addressed later in this section.

Consider Fig. 9.11, and call the band extrema states as $|n\mathbf{k}\rangle = |v\rangle$ and $|m\mathbf{k}\rangle = |c\rangle$. Let us denote $p_{cv} = \langle c|\hat{p}|v\rangle$. If we consider the fact that the **k**- states near the bandgap of semiconductors are formed at the Brillouin zone edge where the intersecting bands differ in wavevector by $2\pi/a$, the magnitude of the momentum matrix element should be $p_{cv} \approx \hbar\frac{2\pi}{a} = \frac{h}{a}$ if we consider the momentum of the free–electron state with this **k** *before* the periodic perturbation. In that case, we get the very useful relation[16]

$$\boxed{\frac{2p_{cv}^2}{m_e} \approx \frac{2h^2}{m_e a^2} \approx 20 \text{ eV},} \tag{9.36}$$

where the numerical value is for a lattice constant $a \sim 5$ Å. Now from Equation 9.35 we can obtain the relation

$$\langle u_c|\nabla_{\mathbf{k}}|u_v\rangle = \frac{\hbar}{E_g}\cdot\frac{p_{cv}}{m_e} \sim \frac{h^2}{2\pi m_e a E_g}, \tag{9.37}$$

which is a characteristic length scale of the periodic part of the Bloch wavefunctions. Exercise 9.6 gives further insight into these matrix elements and their importance in the physics of semiconductors.

Sometimes instead of the momentum matrix elements $\langle mk|\hat{p}|nk\rangle$, we will require to know matrix elements of the real-space coordinate such as $\langle mk|x|nk\rangle$. These will be called the **dipole-matrix** elements, because an elementary dipole is characterized by a charge and a characteristic *length*. To find this, consider the following property of any general system following the Schrödinger equation: $[\hat{H}, x] = \hat{H}x - x\hat{H}$, which with $\hat{H} = \frac{\hat{p}^2}{2m_e} + V$ acting on any function f gives

$$\begin{gathered}
[\frac{\hat{p}^2}{2m_e} + V, x]f = \frac{1}{2m_e}[\hat{p}^2, x]f = \frac{1}{2m_e}(\hat{p}^2(xf) - x\hat{p}^2 f) \\
= \frac{-\hbar^2}{2m_e}(\partial_x(f + xf') - xf'') = \frac{-\hbar^2}{2m_e}(f' + f' + xf'' - xf'') = -\frac{i\hbar}{m_e}\hat{p}f \\
\implies -\frac{i\hbar}{m_e}\hat{p} = [\hat{H}, x], \implies \boxed{-\frac{i\hbar}{m_e}\hat{\mathbf{p}} = [\hat{H}, \mathbf{r}].}
\end{gathered} \tag{9.38}$$

Using this identity, we obtain for Bloch states

$$-\frac{i\hbar}{m_e}\langle\psi_{mk}|\hat{p}|\psi_{nk}\rangle = \langle\psi_{mk}|[\hat{H}, x]|\psi_{nk}\rangle = \underbrace{\langle\psi_{mk}|\hat{H}}_{E_m(k)\langle\psi_{mk}|}\ x - x\ \underbrace{\hat{H}|\psi_{nk}\rangle}_{E_n(k)|\psi_{nk}\rangle}$$

$$\implies \boxed{\langle\psi_{m\mathbf{k}}|\mathbf{p}|\psi_{n\mathbf{k}}\rangle = im_e(\frac{E_m(\mathbf{k}) - E_n(\mathbf{k})}{\hbar})\langle\psi_{m\mathbf{k}}|\mathbf{r}|u_{n\mathbf{k}}\rangle.} \tag{9.39}$$

Note that the right side of the boxed relation for the momentum matrix element is dimensionally in the form $\sim im_e\omega x \sim m_e \times$ (velocity).

---

[16] The combination $2p_{cv}^2/m_e$ appears in various bandstructure calculations and photonic matrix elements, and does not change much from one semiconductor to another. It is useful to remember its magnitude is of the order of 20 eV

Using Equation 9.35, we get

$$\underbrace{\langle \psi_{mk}|\hat{H}}_{E_m(k)\langle\psi_{mk}|}\ x - x\ \underbrace{\hat{H}|\psi_{nk}\rangle}_{E_n(k)|\psi_{nk}\rangle} = i(E_m(\mathbf{k}) - E_n(\mathbf{k}))\langle u_{mk}|\frac{\partial}{\partial k}|u_{nk}\rangle$$

$$\implies \langle \psi_{mk}|x|\psi_{nk}\rangle = \langle u_{mk}|i\frac{\partial}{\partial k}|u_{nk}\rangle$$

$$\implies \boxed{\langle \psi_{m\mathbf{k}}|\mathbf{r}|\psi_{n\mathbf{k}}\rangle = \langle u_{m\mathbf{k}}|i\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle.} \qquad (9.40)$$

This result means that the dipole matrix element between Bloch states in separate bands but of same **k** is equal to the expectation of the gradient in the **k**-space of the *periodic part* of the Bloch function. We will find in later chapters that at the bottom of conduction bands of semiconductors, the periodic part of the Bloch state, u_c, will be typically composed of s-orbital states, and the periodic part of the top of valence band Bloch states u_v, will be p-orbital states. Electromagnetic radiation, or light will introduce a term proportional to **p** in the problem, and this will give a non-zero matrix element for optical transitions between bands of direct-bandgap semiconductors. We define a quantity closely related to Equation 9.40:

$$\mathcal{A}_{n\mathbf{k}} = \langle \psi_{n\mathbf{k}}|\mathbf{r}|\psi_{n\mathbf{k}}\rangle = \langle u_{n\mathbf{k}}|i\nabla_{\mathbf{k}}|u_{n\mathbf{k}}\rangle = \int_{u.c.} u_{n\mathbf{k}}(i\nabla_{\mathbf{k}})u_{n\mathbf{k}}d^3r, \qquad (9.41)$$

called the **Berry connection** of band n. The integral is limited to a unit cell in real space. The Berry connection has units of length, and is of the form $u_{\mathbf{k}} \cdot \frac{u_{\mathbf{k}+\Delta\mathbf{k}} - u_{\mathbf{k}}}{\Delta\mathbf{k}}$, which measures the change of phase angle of $|u_{\mathbf{k}}\rangle$ with **k**. This is analogous to $\mathbf{a}\cdot\mathbf{b} = ab\cos\theta$. The Berry connection accounts for corrections to the velocity term in the dynamics of Bloch states of crystals that lack spatial inversion symmetry and time-reversal symmetry[17].

9.5 Dynamics of Bloch states

Now the response of a Bloch electron in response to an external force F may be found using semi-classical arguments first, before the full quantum derivation. The electron gains energy

$$dE(k) = F\underbrace{vdt}_{dx} = F\frac{1}{\hbar}\frac{dE(k)}{dk}dt \implies F = \hbar\frac{dk}{dt} \implies \boxed{\mathbf{F} = \hbar\frac{d\mathbf{k}}{dt}.} \qquad (9.42)$$

For the quantum derivation, we ask the following questions:

- If a force F acts on a Bloch eigenstate $\psi_k(x)$ for a time t, what is the new quantum state after the time evolution?
- Is this new state also a Bloch eigenstate?

We will find that the answer to the second question is yes: under a constant force, a Bloch eigenstate of crystal momentum **k** evolves into another Bloch eigenstate, but with a different crystal momentum $\mathbf{k}'$. The total Hamiltonian in the presence of a force F is

---

[17] We will discuss these symmetries later in the book. Spatial inversion symmetry is present if upon inversion of coordinates $x \to -x$, all potentials of the crystal look exactly the same. Time reversal symmetry is present if the potentials look identical upon $t \to -t$, meaning running a video backwards. Time-reversal symmetry is broken in ferromagnetic crystals, because they have an internal magnetic field.

$$H = \frac{p^2}{2m_e} + V + Fx. \tag{9.43}$$

Under the external force F, the initial Bloch eigenstate evolves in time t to the state

$$\psi(x,k,t) = e^{-\frac{i}{\hbar}Ht}\psi(x,k_0,t_0) = e^{-\frac{i}{\hbar}(\frac{p^2}{2m_e}+V+Fx)t}\psi(x,k_0,t_0). \tag{9.44}$$

Is this a Bloch eigenstate? We put it to test by translating $x \to x + a$. If we find $\psi(x+a,k,t) = e^{ik(t)a}\psi(x,k,t)$, then it is a Bloch eigenstate whose crystal momentum is $k(t)$. We get

$$\begin{aligned}\psi(x+a,k,t) &= e^{-\frac{i}{\hbar}(\frac{p^2}{2m_e}+V(x+a)+F(x+a))t}\psi(x+a,k_0,t_0)\\ &= e^{-\frac{i}{\hbar}Fat}e^{ik_0a}e^{-\frac{i}{\hbar}(\frac{p^2}{2m_e}+V(x)+Fx)t}\psi(x,k_0,t_0) = e^{i(k_0-\frac{F}{\hbar}t)a}\psi(x,k_0,t_0),\end{aligned} \tag{9.45}$$

which is a Bloch eigenstate of crystal momentum $k(t) = k_0 - \frac{F}{\hbar}t$. This means the force F has moved the eigenstate from k_0 to $k(t) = k_0 - \frac{F}{\hbar}t$ linearly with time.

The equation governing the dynamics of Bloch states in response to a force $\mathbf{F}$ is then given by

$$\boxed{\mathbf{F} = \hbar\frac{d\mathbf{k}}{dt}.}\ \textbf{F} \text{ is an } \textbf{external} \text{ force, } \textbf{k} \text{ is the } \textbf{crystal momentum}. \tag{9.46}$$

If we assume $\mathbf{p} = \hbar\mathbf{k}$, this equation looks deceptively like Newton's law of classical mechanics, stating that the force is equal to the rate of change of momentum of a classical particle. However, we have gone through the entirety of quantum mechanics, learnt how an electron behaves as a wave, and how its dynamics is very different from what classical physics would predict. But it seems we are back to classical mechanics after all?

This viewpoint is very incorrect – in spite of its appearance, Equation 9.46 is a whole different beast from Newton's law! This is because $\mathbf{F}$ here is *not* the *total* force experienced by an electron in a crystal – it is only the *external* force due to electric or magnetic fields applied from outside the crystal. The electron experiences far larger *internal* forces due to the positively charged atomic nuclei, other electrons, and these forces are periodic in space. Where did these forces go? They are cleverly now hidden in the $\mathbf{k}$-term, which is the crystal momentum and is connected quantum-mechanically to the energy. In a crystal the energy of a $\mathbf{k}$ state, $E(\mathbf{k})$ contains all information about the atomic potential and the periodicity. As the *external* force sweeps out $\mathbf{k}$ linearly with time, the energy of the electron, and it's velocity $\mathbf{v}(\mathbf{k})$ change completely quantum mechanically according to the results derived for the Bloch state. This re-emphasizes that $\hbar\mathbf{k}$ is *not* the total momentum of the electron, but through its connection with $E(\mathbf{k})$ and $\mathbf{v}(\mathbf{k})$, is the crystal momentum, and captures the full quantum mechanics of the electron in the crystal potential in all its glory!

To drive home the difference further, consider an electron in the state $\mathbf{k}$ in a crystal with bandstructure $E(\mathbf{k})$. It has a group velocity given by Equation 9.31, $\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k})$. In the *absence* of external forces, $\mathbf{F} = 0$, and the electron stays in that state – meaning it moves with a velocity $\mathbf{v}(\mathbf{k})$, which in a semiconductor can be of the order of 10^7 cm/s. This may seem like Newton's law all over again – in the absence of a force, an object should stay at rest, or a state of constant velocity – here we seem to have a situation of a constant velocity. But wait, the electron is experiencing very large forces due to the atoms and the periodic potential $V(r+a) = V(r)$. As Fig. 9.12 shows, classically it should be accelerating and decelerating just as a ball rolling over hills and valleys! But the electron maintains a *constant* velocity – this behavior captures the remarkable consequence of quantum mechanics, and the wave-particle duality of the electron.

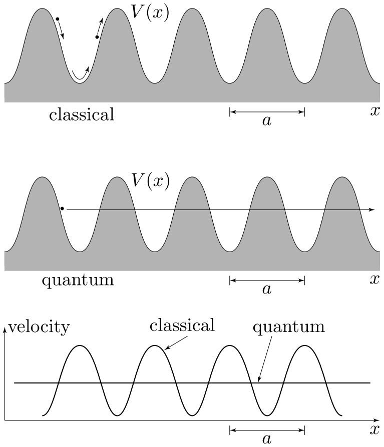


Fig. 9.12 Classical vs. quantum pictures of a particle in a periodic potential. The Bloch state maintains the same velocity in spite of a rapidly varying periodic potential of the crystal – this is impossible in classical mechanics.

Degeneracies in $E_n(\mathbf{k})$ and eigenvector vortices: The Berry phase: As was alluded to earlier, Equation 9.31 for the group velocity $\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k})$ can handle energy bandstructures $E(\mathbf{k})$ that vary smoothly in the $\mathbf{k}$-space. But in some crystals (for example, 2D graphene), energy bands touch and form conical sections at specific points in the $\mathbf{k}$-space, as shown schematically in Fig. 9.13. The group velocity goes to zero at parabolic band extrema, but at conical points, the velocity vector points radially outwards and does not change in magnitude. Therefore, the direction of the vector is ill defined at the point of band intersection. If the eigenvectors are plotted, they will be seen to vary smoothly around parabolic band extrema, like the "laminar" flow of water. But around band intersections, the eigenvectors twirl around, similar to vortices in the flow of water.

The integral of the eigenvectors around the point of degeneracy results in a phase, called the Berry phase (see Fig. 9.14). The group velocity has to be generalized to account for such eigenvector singularities in the $\mathbf{k}$-space. The modified expression is

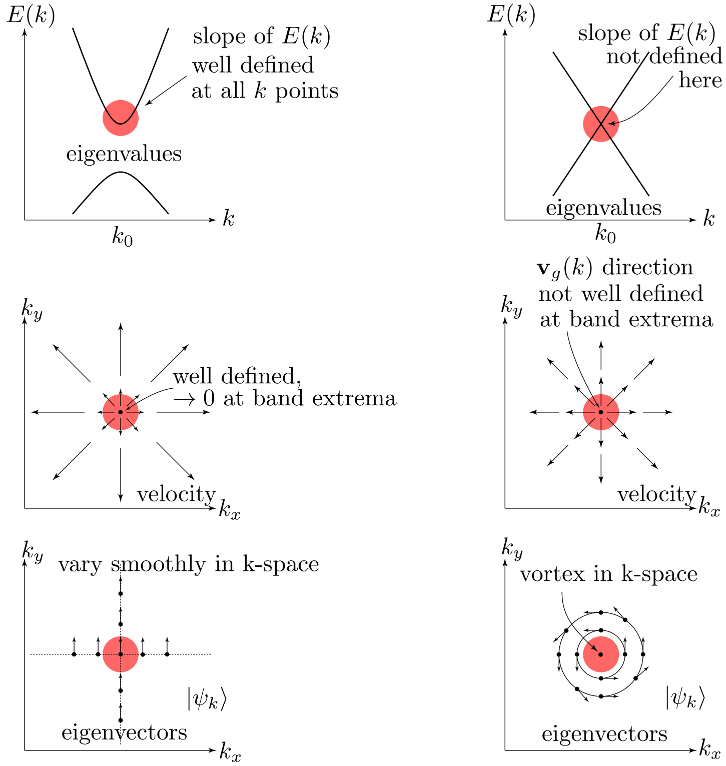
$$\boxed{\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k}) - \frac{\mathbf{F}}{\hbar} \times \mathbf{\Omega}_{\mathbf{k}}.} \tag{9.47}$$

Fig. 9.14 Sir Michael Berry, mathematical physicist who in 1983 unearthed the Berry phase in quantum mechanics. This geometric phase was hiding in plain sight for half a century; its consequences are physically observable.

where $\mathbf{F} = \hbar\frac{d\mathbf{k}}{dt}$ is the external force on the electron, and

$$\begin{aligned}\mathbf{\Omega}_{\mathbf{k}} = \nabla_{\mathbf{k}} \times \mathcal{A}_{\mathbf{k}} = \nabla_{\mathbf{k}} \times \int_{u.c.} u_{n\mathbf{k}}(i\nabla_{\mathbf{k}})u_{n\mathbf{k}}d^3r \\ = \nabla_{\mathbf{k}} \times \langle u_{\mathbf{k}}|i\nabla_{\mathbf{k}}|u_{\mathbf{k}}\rangle = i\langle\nabla_{\mathbf{k}}u_{\mathbf{k}}| \times |\nabla_{\mathbf{k}}u_{\mathbf{k}}\rangle\end{aligned} \tag{9.48}$$

is the **Berry curvature** of a band. $\mathcal{A}_{\mathbf{k}}$ is the Berry connection defined in Equation 9.41 (Exercise 9.7 outlines a derivation). The Berry curvature plays the role of a quasi-magnetic field originating internally, from the symmetries inherent to the crystal. In analogy to the vector potential


$E(k)$
slope of $E(k)$
well defined
at all k points
eigenvalues
k_0
k
$E(k)$
slope of $E(k)$
not defined
here
eigenvalues
k_0
k
$\mathbf{v}_g(k)$ direction
not well defined
at band extrema
k_y
well defined,
$\rightarrow 0$ at band extrema
velocity
k_x
k_y
velocity
k_x
k_y
vary smoothly in k-space
$|\psi_k\rangle$
eigenvectors
k_x
k_y
vortex in k-space
$|\psi_k\rangle$
eigenvectors
k_x

$$\mathbf{v}_g(k) = \frac{1}{\hbar}\nabla_k E(k) \rightarrow \boxed{\mathbf{v}_g(k) = \frac{1}{\hbar}\nabla_k E(k) + \frac{d\mathbf{k}}{dt} \times \underbrace{\nabla_k \times \langle u_k | i\nabla_k | u_k \rangle}_{\Omega(k),\ \text{Berry curvature}}}$$

Fig. 9.13 Berry vortices.

$\mathbf{A}$ of a real magnetic field $\mathbf{B} = \nabla_{\mathbf{r}} \times \mathbf{A}$ due to currents in wires, the quasi-magnetic field originating in $\mathbf{k}$-space follows the relation $\mathbf{\Omega}_{\mathbf{k}} = \nabla_{\mathbf{k}} \times \mathcal{A}_{\mathbf{k}}$. In other words, the Berry curvature is the "curl" of the Berry connection. Equation 9.47 states that in response to an external force $\mathbf{F}$, a velocity *transverse* to the direction of the force[18] is experienced by electrons in Bloch eigenstates in crystals that have a non-zero $\mathbf{\Omega}$. Though the idea of the Berry curvature of Bloch states was motivated by considering situations where the slope of $E(\mathbf{k})$ is not well defined, it is a far more general concept. Non-zero Berry curvatures can be present for crystals with smoothly changing $E(\mathbf{k})$.

For the actual calculation of the Berry curvature, the following formulation due to Michael Berry is convenient. Since for $m \neq n$, the relation $\langle u_{n\mathbf{k}} | \hat{H}_{\mathbf{k}} | u_{m\mathbf{k}} \rangle = E_{m\mathbf{k}} \langle u_{n\mathbf{k}} | u_{m\mathbf{k}} \rangle = 0$ holds, taking $\nabla_{\mathbf{k}}$ on both sides and using Equation 9.35 gives

$$\mathbf{\Omega}_{n\mathbf{k}} = i \sum_{m \neq n} \frac{\langle u_{n\mathbf{k}} | \nabla_{\mathbf{k}} \hat{H}_{\mathbf{k}} | u_{m\mathbf{k}} \rangle \times \langle u_{m\mathbf{k}} | \nabla_{\mathbf{k}} \hat{H}_{\mathbf{k}} | u_{n\mathbf{k}} \rangle}{(E_n(\mathbf{k}) - E_m(\mathbf{k}))^2}, \tag{9.49}$$

which on using Equations 9.40 and 9.35 becomes

---

[18] The transverse force leads to the *anomalous* Hall effect observed in magnetic crystals. With the Berry curvature term, the phenomena is not anomalous any more!

$$\boxed{\mathbf{\Omega}_{n\mathbf{k}} = i\hbar^2 \sum_{m\neq n} \frac{\langle \psi_{n\mathbf{k}}|\hat{\mathbf{v}}|\psi_{m\mathbf{k}}\rangle \times \langle \psi_{m\mathbf{k}}|\hat{\mathbf{v}}|\psi_{n\mathbf{k}}\rangle}{(E_n(\mathbf{k}) - E_m(\mathbf{k}))^2}.} \tag{9.50}$$

The Berry curvature $\mathbf{\Omega}$ has units of (length)^2, or area, and the Berry connection $\mathcal{A}$ has units of length; they are related by $\mathbf{\Omega} = \nabla_{\mathbf{k}} \times \mathcal{A}$. Table 9.2 shows which symmetries of the crystal dictate the presence, or absence of a Berry curvature. Note that the $\mathbf{m}_{\mathbf{k}}$ is the orbital angular momentum of a Bloch wavepacket, a concept that we have not discussed in this book yet. If a crystal has both time-reversal symmetry (e.g. it is non-magnetic), and has inversion symmetry (e.g. in a monoatomic crystal), then the Berry curvature has to satisfy $\mathbf{\Omega}(\mathbf{k}) = -\mathbf{\Omega}(-\mathbf{k}) = \mathbf{\Omega}(-\mathbf{k})$. Since a number can be equal to its negative only if it is zero, **the Berry curvature is zero for a crystal that has both time reversal and inversion symmetries**. We will encounter the Berry phase and its consequences later in this book.

Table 9.2 Symmetry dictates the presence or absence of Berry curvature in the bandstructure of crystalline solids.

	Time Reversal	Inversion
action	$t \to -t$	$x \to -x$
flip sign		
	$\mathbf{k}$	$\mathbf{r}$
	$-i\nabla_{\mathbf{r}}$	$\mathbf{k}$
	$\frac{d\mathbf{r}}{dt}$	$\mathcal{A}_{\mathbf{k}}$
	$\mathbf{B}$	$\mathbf{E}$
don't flip		
	$\mathbf{E}$	$\mathbf{B}$
	$\mathcal{A}_{\mathbf{k}}$	
result		
	$\mathbf{\Omega}_{\mathbf{k}} = -\mathbf{\Omega}_{-\mathbf{k}}$	$\mathbf{\Omega}_{\mathbf{k}} = \mathbf{\Omega}_{-\mathbf{k}}$
	$\mathbf{m}_{\mathbf{k}} = -\mathbf{m}_{-\mathbf{k}}$	$\mathbf{m}_{\mathbf{k}} = \mathbf{m}_{-\mathbf{k}}$

9.6 Bloch wave velocity and ballistic current

In earlier chapters, we evaluated the quantum mechanical current carried by a distribution of electron states by summing the contributions from individual occupied states. In Equation 9.31 we derived that corresponding to the quantum mechanical velocity of a free–electron $\psi(x) = \frac{1}{\sqrt{L}}e^{ikx}$ as $v(k) = \frac{\hbar k}{m}$, the velocity of a Bloch state simply changes to the ”group velocity” when the electron is put in a periodic crystal with Block eigenfunction $\psi(x) = \frac{1}{\sqrt{L}}e^{ikx}u(x)$. The group velocity of Bloch state $\mathbf{k}$ is simply $\mathbf{v}_g(\mathbf{k}) = \nabla_{\mathbf{k}}E(\mathbf{k})/\hbar$, which is known when the electron bandstructure is known – *without recourse to the wavefunction*. Sometimes, we trade the knowledge of wavefunction for the knowledge of the bandstructure $E(\mathbf{k})$, which is attractive because the bandstructure of semiconductors are experimentally measured and tabulated. We also generalize to any dimension d. The expression for the quantum current carried by various Bloch states then becomes

$$\mathbf{J}_d = \frac{q}{L^d}\sum_{\mathbf{k}} \mathbf{v}_g(\mathbf{k})f(\mathbf{k}), \tag{9.51}$$

where we have now included the occupation probability $f(\mathbf{k})$ of state $|\mathbf{k}\rangle$. We next split off the spin degeneracy $g_s = 2$, and allow a valley degeneracy g_v for each $\mathbf{k}$-state. Picture the current flow from a ”left” contact to a ”right” contact. Let us also allow for scattering, which are quantum-mechanical reflections in going from the left to the right contact. Let the quantum-mechanical transmission coefficient from the left to the right contact be $T(\mathbf{k})$. Then, the expression for the net current density flowing from the left to the right becomes

$$\mathbf{J}_d = \frac{qg_sg_v}{L^d}\sum_{\mathbf{k}} \mathbf{v}_g(\mathbf{k})T(\mathbf{k})[f_L(\mathbf{k}) - f_R(\mathbf{k})]. \tag{9.52}$$

The occupation function difference results because $f_L(\mathbf{k})[1-f_R(\mathbf{k})] - f_R(\mathbf{k})[1-f_L(\mathbf{k})] = f_L(\mathbf{k}) - f_R(\mathbf{k})$. The transmission coefficient $T(\mathbf{k})$ does not depend on the direction of current flow.

The sum over **k** states is typically always converted into an integral. The recipe for this step uses the fact that the allowed states in the **k**-space are discrete "boxes" of each side $2\pi/L$ and volume $(2\pi/L)^d$ in d dimensions. Then, the sum converts to an integral via

$$\sum_{\mathbf{k}}(...) = \int \frac{d^d\mathbf{k}}{(\frac{2\pi}{L})^d}(...). \tag{9.53}$$

The expression for the current density is then[19]

$$\mathbf{J}_d = \frac{q g_s g_v}{(2\pi)^d}\int d^d\mathbf{k} \times \mathbf{v}_g(\mathbf{k})T(\mathbf{k})[f_L(\mathbf{k}) - f_R(\mathbf{k})], \tag{9.54}$$

which using the velocity of Bloch states becomes

$$\mathbf{J}_d = \frac{q g_s g_v}{\hbar(2\pi)^d}\int d^d\mathbf{k}[\nabla_{\mathbf{k}}E(\mathbf{k})]T(\mathbf{k})[f_L(\mathbf{k}) - f_R(\mathbf{k})]. \tag{9.55}$$

The unit of current density is in $\mathrm{A/m}^{d-1}$ for current flow in d-dimensions. This expression for the current density is applicable in a wide range of situations ranging from ballistic transport, to scattering limited drift or diffusion, and to tunneling transport, and in multiple dimensions. The group velocity term $\mathbf{v}_g(\mathbf{k})$ is locked down by the electron band structure $E(\mathbf{k})$, and external forces modify the occupation functions. For example, the response of $f(\mathbf{k})$ to electric fields, or concentration gradients are tracked by a Boltzmann transport equation.

The group velocity term is a first-order derivative in k. Therefore, for the particular case of $d = 1$ dimension, the current simplifies to

$$I = \frac{q g_s g_v}{\hbar(2\pi)}\int dE \cdot T(E)[f_L(E) - f_R(E)] \approx (g_s g_v \frac{q^2}{h})V \tag{9.56}$$

if we assume $T(E) = 1$ for perfect (ballistic) transmission and since $\int dE[f_L(E) - f_R(E)] = qV$ at $T \to 0$ K. The ballistic conductance of a 1D mode is therefore given by $G_0 = g_s g_v \times q^2/h$ and does not depend on the exact bandstructure $E(\mathbf{k})$. This independence is the reason the 1D ballistic conductance for a general $E(\mathbf{k})$ is identical to what was found for the free electron derived in Chapter 5, Equation 5.24.

9.7 Transport by Bloch waves with scattering

Because Bloch states are eigenstates for the periodic potential[20], an electron in state **k** is a stationary state: it will remain in that state unless acted upon by an external force. If the periodicity of the crystal potential is broken (e.g. due to defects or impurities), the electron can scatter out of state $\mathbf{k} \to \mathbf{k}'$. In the absence of scattering, the electron

[19] Note the cancellation of the dependence on the macroscopic dimension L. The **k**-coordinate system is chosen to be either cartesian, cylindrical, or spherical based on the specific problem. Also note that the mass of electrons does not appear explicitly: it is hidden inside the group velocity. As a result, this expression for current also works for "massless" energy dispersions, such as for conical bands of 2D graphene.

[20] In real semiconductor devices, the techniques of doping and quantized heterostructures introduce electron wavefunctions that are not necessarily pure Bloch states. Since the probability density of Bloch states $|\psi_{\mathbf{k}}(\mathbf{r})|^2$ is periodic in the entire crystal, this is often not desirable for device applications. In such cases, linear combinations of various Bloch **k** states, called a *wavepacket* captures the real electronic states, and we track the equations of motion of the wavepackets instead of the pure Bloch states. This is the subject of discussion of Chapters 13 and subsequent chapters on electron transport and photonic properties. What we learn here for pure Bloch states carries over without change to the Bloch wavepackets, the theory of which goes under the name of the *effective mass approximation*.

wavevector $\mathbf{k}$ upon the application of an *external* force $\mathbf{F}$ changes with time as $\mathbf{F} = \hbar \frac{d\mathbf{k}}{dt}$, where $\hbar\mathbf{k}$ is the *crystal* momentum. Consider the case when the electron was in equilibrium in state $\mathbf{k}(0)$ before a force $\mathbf{F}$ was applied to it. Just as in the Drude model, if after τ seconds, the electron state scatters to state $\mathbf{k}'$, the equation of motion changes to

$$\mathbf{F} = \hbar \frac{d\mathbf{k}}{dt} - \frac{\hbar \Delta \mathbf{k}}{\tau_{\mathbf{k}}}, \tag{9.57}$$

where the second term on the right is the "friction" due to scattering, with $\Delta\mathbf{k} = \mathbf{k}(t) - \mathbf{k}(0)$. In the steady state $\frac{d}{dt}(...) \to 0$, and the electron will occupy the $\mathbf{k}$ state given by $\mathbf{k}(t) = \mathbf{k}(0) - \frac{\mathbf{F}\tau_{\mathbf{k}}}{\hbar}$. Another way to understand this is that if the electron started in state $\mathbf{k}(0)$, it will end up in state $\mathbf{k}(0) - \frac{\mathbf{F}\tau_{\mathbf{k}}}{\hbar}$ after a long time, when steady state is reached. Now this statement is true for *all* $\mathbf{k}$-states! How exactly the occupation function $f(\mathbf{k})$ changes in response to external forces, and internal scattering processes, is the topic of discussion in Chapter 19: it is given by the **Boltzmann transport equation**[21]. Here, we discuss the simplest possibility: that the entire electron occupation function $f(\mathbf{k})$ rigidly shifts to $f(\mathbf{k}(0) - \frac{\mathbf{F}\tau_{\mathbf{k}}}{\hbar})$ without changing its shape. Then, the quantum mechanical current in the presence of scattering is given by

$$J = q\frac{g_s g_v}{L}\sum_k v_g(k) f(k) = q\frac{g_s g_v}{L}\sum_k v_g(k) f(\mathbf{k}(0) - \frac{\mathbf{F}\tau_{\mathbf{k}}}{\hbar}). \tag{9.58}$$

For small forces $\mathbf{F}$, we may Taylor-expand the distribution function $f(\mathbf{k}(0) - \frac{\mathbf{F}\tau_{\mathbf{k}}}{\hbar}) \approx f(\mathbf{k}(0)) - \frac{\mathbf{F}\tau_{\mathbf{k}}}{\hbar} \cdot \nabla_{\mathbf{k}} f(\mathbf{k}(0))$, and obtain the current density as

$$J = q\frac{g_s g_v}{L}\sum_k v_g(k) f(\mathbf{k}(0) - \frac{\mathbf{F}\tau_{\mathbf{k}}}{\hbar})$$

$$\approx \underbrace{q\frac{g_s g_v}{L}\sum_k v_g(k) f(\mathbf{k}(0))}_{0} + q\frac{g_s g_v}{L}\frac{\mathbf{F}}{\hbar}\cdot\sum_k v_g(k)\tau_{\mathbf{k}}(-\nabla_{\mathbf{k}} f(\mathbf{k}(0))), \tag{9.59}$$

where the first term on the bottom row is zero because the summation is performed over the equilibrium occupation function $f(\mathbf{k}(0))$, which is *symmetric* in $\mathbf{k}$ because it is a Fermi–Dirac distribution function (Fig. 9.15). The general form of the current given by the remaining asymmetric component of the distribution function is picked up by the derivative $\nabla_{\mathbf{k}} f(...)$, discussed in detail in Chapter 21. Since the total electron density in d-dimensions is given by

$$n = \frac{g_s g_v}{L^d}\sum_k f(k), \tag{9.60}$$

if the bandstructure is parabolic: $E(\mathbf{k}) = \frac{\hbar^2|\mathbf{k}|^2}{2m^\star}$, and the external force is due to a constant electric field, $\mathbf{F} = q\mathbf{E}$, the magnitude of the current

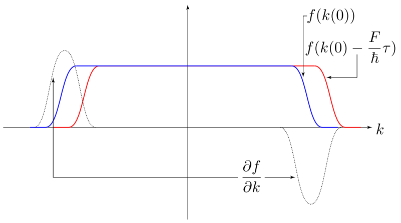

$f(k(0))$
$f(k(0) - \frac{F}{\hbar}\tau)$
k
$\frac{\partial f}{\partial k}$

Fig. 9.15 The occupation function shifts from the equilibrium Fermi–Dirac value upon the application of an external force. The sketch indicates $f(\mathbf{k})$, $f(\mathbf{k} - \frac{\mathbf{F}}{\hbar}\tau)$, and $\nabla_{\mathbf{k}} f(\mathbf{k})$.

[21] The solution of the Boltzmann equation is derived and discussed in Chapter 19. In response to gradients in electric potential V and temperature T, the solution in the linear approximation is $g_k \approx f_0 + (-\frac{\partial f_0}{\partial E_k})\tau_k \mathbf{v}(\mathbf{k}) \cdot (-q\nabla V - \frac{E_k - E_F}{T}\nabla T)$, where f_0 is the equilibrium distribution function and the second term is the perturbation caused by the electric field and temperature gradients.

density carried by Bloch states in d-dimensions is

$$J \approx \frac{nq^2}{m^\star} \underbrace{\left(\frac{2}{d}\frac{\sum_k E_{\mathbf{k}}\tau_{\mathbf{k}}(-\frac{\partial f(k)}{\partial E_{\mathbf{k}}})}{\sum_k f(k)}\right)}_{\langle\tau\rangle} E = \underbrace{\frac{nq^2\langle\tau\rangle}{m^\star}}_{\sigma} E = qn \underbrace{\frac{q\langle\tau\rangle}{m^\star}}_{\mu} E, \quad (9.61)$$

where $\langle\tau\rangle$ is an effective scattering time that is a weighted, and ensemble averaged time constant over the Bloch $\mathbf{k}$ states[22]. Let us compare the conductivity and mobility of Bloch states to the classical Drude-model of Chapter 2 for electron conductivity in Equation 2.3 $\sigma = \frac{nq^2\tau}{m_e}$, and electron mobility $\mu = \frac{q\tau}{m_e}$. The updated results for the Bloch states are the electronic conductivity $\sigma = \frac{nq^2\langle\tau\rangle}{m^\star}$ and mobility $\mu = \frac{q\langle\tau\rangle}{m^\star}$. Unlike the classical Drude model, the free–electron mass m_e is replaced by the effective mass of the energy band. These are quantum mechanical results, because the electron bandstructure of Bloch states enters via $m^\star$, and the individual scattering rates from each Bloch state $\mathbf{k} \rightarrow \mathbf{k}'$ via the Fermi golden rule. The individual scattering rates between Bloch states is the topic of discussion in Chapter 20. One typically still refers to the new formulation for transport of Bloch states with scattering as an effective Drude model, but one must use the correct effective masses and scattering times to compare to experiments.

9.8 Energy (heat) current

The flow of an electron of energy $E(\mathbf{k})$ transports not just charge, but the associated energy as well. The energy current density is given by $\mathbf{J}_E(\mathbf{k}) = \sum_{\mathbf{k}} E(\mathbf{k}) \cdot \mathbf{v}(\mathbf{k})/L^d$, and thus we obtain

$$\mathbf{J}_E = \frac{g_s g_v}{(2\pi)^d} \int d^d\mathbf{k} E(\mathbf{k})\mathbf{v}(\mathbf{k})T(\mathbf{k})[f_L(\mathbf{k}) - f_R(\mathbf{k})]. \quad (9.62)$$

The units are in watts/m^{d-1} for transport in d-dimensions. Using the same approximations as earlier for a 1D conductor, we obtain the energy current carried by electrons to be $I_E \approx (g_s g_v q^2/h) \cdot V^2$ in Watts if the transport is ballistic ($T(\mathbf{k}) = 1$) at $T \rightarrow 0$ K.

In the presence of scattering, the heat current density drives the distribution function to a non-equilibrium value due to a temperature gradient, $\nabla_{\mathbf{r}} T(\mathbf{r})$, or $\frac{dT(x)}{dx}$ in one dimension. In Chapter 21, we will find from the Boltzmann transport equation that a temperature gradient drives the equilibrium distribution function $f(\mathbf{k}(0))$ to $f(\mathbf{k}) \approx f(\mathbf{k}(0)) - \tau_{\mathbf{k}}\mathbf{v}(\mathbf{k})k_b \ln(\frac{N_c}{n})\frac{\partial f(\mathbf{k})}{\partial E(\mathbf{k})}\nabla_{\mathbf{r}} T(\mathbf{r})$, where n is the electron concentration, and N_c is the band-edge DOS. This leads to a heat current density

$$\mathbf{J}_E = \frac{g_s g_v}{(2\pi)^d} \int d^d\mathbf{k}[E(\mathbf{k}) - E_F]\mathbf{v}(\mathbf{k})[f(\mathbf{k}(0)) - \tau_{\mathbf{k}}\mathbf{v}(\mathbf{k})k_b \ln(\frac{N_c}{n})\frac{\partial f(\mathbf{k})}{\partial E(\mathbf{k})}\nabla_{\mathbf{r}} T(\mathbf{r})]. \quad (9.63)$$

---

[22] Note that the expression for the current density for Bloch states now looks deceptively similar to the classical Drude model! However, this expression captures the quantum mechanical behavior of electrons because of the effective mass, and the details of the scattering time $\langle\tau\rangle$.

Similar to the charge current, the first term cancels because $\mathbf{v}(\mathbf{k})$ is odd in $\mathbf{k}$ whereas $E(\mathbf{k})$ and $f(\mathbf{k}(0))$ are even. The second term gives a heat current density

$$\mathbf{J}_E = \underbrace{\frac{g_s g_v}{(2\pi)^d} \ln(\frac{N_c}{n}) \Big[k_b \int d^d\mathbf{k} [E(\mathbf{k}) - E_F] \mathbf{v}^2(\mathbf{k}) \tau_{\mathbf{k}} (-\frac{\partial f(\mathbf{k})}{\partial E(\mathbf{k})}) \Big]}_{\kappa} \cdot \nabla_{\mathbf{r}} T(\mathbf{r}), \tag{9.64}$$

where the thermal conductivity κ is now related to the bandstructure of the Bloch states via the group velocity $\mathbf{v}(\mathbf{k})$, the energy eigenvalues $E(\mathbf{k})$, and the scattering times $\tau_{\mathbf{k}}$. The integral above is typically easily evaluated with the knowledge of the electron bandstructure $E(\mathbf{k})$ and the corresponding group velocities $\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar} \nabla_{\mathbf{k}} E(\mathbf{k})$, and the occupation function either in the Fermi–Dirac form, or in the simpler Maxwell–Boltzmann form. Chapter 21 discusses several aspects of heat current in more detail.

9.9 Any current

Following the above recipes, spin currents, polarization currents, and other related quantities may be evaluated. We realize from the results derived in this chapter that the key to finding the current are the occupation functions $f(\mathbf{k})$ and their dependence on external voltages. This indeed is the driver of current for all cases. Determination of the occupation functions, and their dependence on external voltages is the topic of Chapter 21. As a simple example, the net spin current density can be written as the difference of the two spin components, where the electron carries with it a spin angular momentum of $\frac{\hbar}{2}$:

$$\mathbf{J}_{spin} = \underbrace{\frac{\hbar}{2} \frac{g_v}{L^d} \sum_{\mathbf{k},\uparrow} \mathbf{v}(\mathbf{k},\uparrow) f(\mathbf{k},\uparrow)}_{\mathbf{J}_\uparrow} - \underbrace{\frac{\hbar}{2} \frac{g_v}{L^d} \sum_{\mathbf{k},\downarrow} \mathbf{v}(\mathbf{k},\downarrow) f(\mathbf{k},\downarrow)}_{\mathbf{J}_\downarrow}, \tag{9.65}$$

whereas the corresponding charge current is the *sum* of the two components multiplied by the electron charge:

$$\mathbf{J}_{charge} = \underbrace{q \frac{g_v}{L^d} \sum_{\mathbf{k},\uparrow} \mathbf{v}(\mathbf{k},\uparrow) f(\mathbf{k},\uparrow)}_{\mathbf{J}_\uparrow} + \underbrace{q \frac{g_v}{L^d} \sum_{\mathbf{k},\downarrow} \mathbf{v}(\mathbf{k},\downarrow) f(\mathbf{k},\downarrow)}_{\mathbf{J}_\downarrow}. \tag{9.66}$$

In some devices such as spin-torque transfer memories, or magnetic semiconductors, it may be useful to ask how much spin current is carried by a charge current, or what is the ”charge to spin conversion”. This is found as the ratio of the two quantities above:

$$\alpha = \frac{\mathbf{J}_{spin}}{\mathbf{J}_{charge}} = \frac{\hbar}{2q} \cdot \frac{\sum_{\mathbf{k},\uparrow} \mathbf{v}(\mathbf{k},\uparrow) f(\mathbf{k},\uparrow) - \sum_{\mathbf{k},\downarrow} \mathbf{v}(\mathbf{k},\downarrow) f(\mathbf{k},\downarrow)}{\sum_{\mathbf{k},\uparrow} \mathbf{v}(\mathbf{k},\uparrow) f(\mathbf{k},\uparrow) + \sum_{\mathbf{k},\downarrow} \mathbf{v}(\mathbf{k},\downarrow) f(\mathbf{k},\downarrow)}). \tag{9.67}$$

If each **k**-state has both up and down spin states occupied, then the spin current is zero, and the charge-to-spin conversion $\alpha = 0$. On the other hand, if all spins are polarized in the same direction, say up, the maximum charge to spin conversion is $\hbar/2q$, in which each electron transports $\hbar/2$ angular momentum of spin, with no cancellation with oppositely moving electrons.

9.10 Quantum Wiedemann–Franz law

The ratio of the thermal and electronic conductivities for electron Bloch states is the quantum Lorenz number:

$$\mathcal{L}_q = \frac{\kappa}{\sigma T} = \frac{g_s g_v \ln(\frac{N_c}{n}) k_b \int \frac{d^dk}{(2\pi)^d} v^2 \tau E \frac{\partial f}{\partial E}}{T \cdot q^2 \cdot g_s g_v \int \frac{d^dk}{(2\pi)^d} v^2 \tau \frac{\partial f}{\partial E}} \underbrace{\rightarrow}_{\text{semi-classical approx.}} = \frac{\pi^2}{3}(\frac{k_b}{q})^2. \tag{9.68}$$

For metals, the conceptual changes due to the Bloch theorem insert far more details of the bandstructure into this ratio, and resolves the low electronic specific heats. Nevertheless, the ratio is close to the classical result of Chapter 2. However, for semiconductors this relation fails because the net thermal conductivity of a crystal $\kappa = \kappa_{el} + \kappa_{lattice}$ is the sum of the electronic and the lattice contributions. The original Drude model only includes the electronic component to explain the Wiedemann–Franz law. But in semiconductors, and in general electrically non-metallic crystals, heat is primarily carried by lattice vibrations (phonons) and not by electrons, which is why the Wiedemann–Franz law is not valid for them[23].

9.11 Metals, semiconductors, semimetals and insulators

The lattice geometry of a crystal leads to the formation of sharply defined Brillouin zones. The energy eigenvalues allowed for **k**-states within each Brillouin zone form bands, which may or may not be separated by gaps. The valence electrons fill up these bands – if a band is partially filled, it is a metal, as shown in Fig. 9.16. The density of states at the Fermi level $g(E_F)$ is very large for a metal, of the order of $10^{23}/\text{cm}^3$. If the Fermi level falls within an energy gap or very near the gap within the band edges, it is a semiconductor: $0 \leq g(E_F) \lesssim 10^{20}/\text{cm}^3$. If the Fermi level falls within overlapping bands, we have a semimetal, which can host both electrons and holes at the Fermi level. For typical semimetals $g(E_F) \sim 10^{20}/\text{cm}^3$.

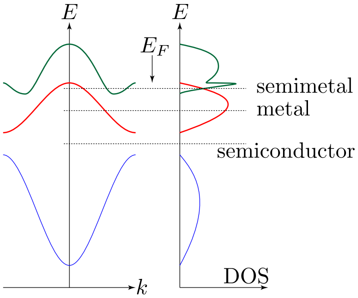


Fig. 9.16 Semiconductors, metals, and semimetals differ primarily by the density of states at their Fermi level, $g(E_F)$. The Fermi level is determined by counting the total valence electrons/basis of a crystal and using the rule that each filled band holds $2N$ electrons.

9.12 Chapter summary section

In this chapter, we learned:

[23] If a thermal gradient can drive a voltage difference between two terminals, we have a *thermal battery* that converts heat energy to electrical energy. The opposite process, of controlling the temperature by voltage differences is also possible with such materials, which are called **thermoelectrics**. Semiconductors and semimetals are the most efficient thermoelectric materials, because they typically have low thermal conductivities than metals, and high electrical conductivities than insulators. The figure of merit for thermoelectrics is $ZT = S^2 T \frac{\sigma}{\kappa}$, where S is the Seebeck coefficient – because of the Wiedemann–Franz law, this would simply be $ZT = (\frac{Sq}{k_b})^2$ for a metal. But for semiconductors, and especially for semimetals, this can be quite different and large, which makes them efficient thermoelectrics: this is discussed further in Chapters 21 and 22.

- The wavefunctions allowed for electrons in a periodic potential $V(x+a) = V(x)$ of a crystal lattice are Bloch functions of the form $\psi(x) = e^{ikx}u_k(x)$. Though the Bloch wavefunction itself is not periodic with the lattice, the part $u_k(x)$ is, and so is $|\psi(x)|^2$. The energy eigenvalues for each allowed k-state forms bands $E(k)$. They are collectively called the bandstructure, which contain in them all electronic properties of any crystal.
- The substitution of the Bloch eigenfunctions into the Schrödinger equation leads to several recipes for actually calculating the bandstructure. These are the nearly free–electron bandstructure, the tight-binding model, the $\mathbf{k} \cdot \mathbf{p}$ model, the empirical pseudopotential method, and density functional theory models.
- 3D crystals are conceptually split into a basis of atoms that repeats in every cell, and a lattice that contains the periodicity and symmetries. The set of lattice points in real space can be represented in the **k**-space, called the reciprocal space lattice. The smallest repeating volume of the lattice in the real space is called the Wigner–Seitz cell, and the corresponding volume in the **k**-space is called the Brillouin zone. Each BZ contains as many **k**-points as there are Wigner–Seitz cells in real space.
- The energy eigenvalues belonging to each BZ form bands. Each band can hold 2N electrons where N is the number of Wigner–Seitz cells in the entire crystal. The bands of a BZ may or may not overlap with bands of the next BZ. If they do not overlap, there are energy bandgaps between the bands. The Pauli exclusion principle allows two electrons/state, and 2N electrons/band. The highest energy filled is the Fermi level. If the Fermi level is deep within a band, it is a metal. If the Fermi level is inside an energy gap, it is a semiconductor or an insulator. If the Fermi level is at the intersection of two bands, it is a semimetal.
- Bloch states are energy eigenstates, and hence stationary states. The velocity of a Bloch state is the group velocity of that state, obtained from the energy bandstructure as $\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}}E(\mathbf{k})$. Unlike free–electrons, Bloch states are states of definite energy, but *not* of definite momentum. The quantity $\hbar\mathbf{k}$ is called the crystal momentum, and has contributions of the real momentum of the electron, as well as from the crystal potential.
- The charge current carried by electrons in a crystal of d-dimensions is given by $\mathbf{J} = \frac{q}{L^d}\sum_{\mathbf{k}} \mathbf{v}(\mathbf{k})f(\mathbf{k})$, where $f(\mathbf{k})$ is the occupation probability of state $\mathbf{k}$. The sum must be carried out over all valleys and spins as required by the energy bandstructure $E(\mathbf{k})$. Whether the transport is ballistic or limited by scattering does not change the group velocities, as they are locked down by the bandstructure. They only change the occupation function $f(\mathbf{k})$.
- The Bloch theorem and the quantum facelift of the electron and heat transport models changes the Wiedemann–Franz law by explaining why it fails for semiconductors and insulators. In turn, it seeds new ideas for applications, such as thermoelectrics.

Further reading

The books referred to in Chapter 8 are excellent sources for conceptual developments till 1980s, especially those by Mott and Jones, John Ziman, and Ashcroft and Mermin. The popular text **Introduction to Solid State Physics** by Charles Kittel has a good introductory treatment of Bloch function and the nearly free–electron bandstructure.

For concepts that were developed in the field since the 1980s, such as Berry curvature effects on Bloch functions, I recommend two modern texts: **Condensed Matter Physics** by Michael Marder is one of the early texts on the subject that introduces the Berry phase and discusses some of its effects on electronic and magnetic properties of Bloch wavepackets. **Solid State Physics** by Giuseppe Grosso and Giuseppe Pastori Parravicini goes deep into this subject by dedicating a full chapter that connects geometric phases with several physical phenomena of crystals, such as electronic polarization.

Exercises

(9.1) **Summing N roots of unity**

(a) The real roots of 1 are ± 1. But if complex numbers are allowed, show that the N^{th} roots of unity lie on the unit circle on the complex plane at angles $\theta_n = n \times \frac{2\pi}{N}$.

(b) Write down all five of the fifth roots of unity. What is their sum?

(c) Prove that the sum of the N roots is always zero: $\sum_{n=1}^{N} e^{i\theta_n} = 0$.

(9.2) **Miscellaneous**

Give short answers to the following questions:

(a) The Sommerfeld model introduced for the first time the then-newly discovered Fermi–Dirac distribution to explain the conductivity due to free–electrons in solids. Though this model could explain certain thermodynamic properties such as specific heats that the classical Drude model failed to explain, it could not explain why crystals could be metals, insulators, or semiconductors. What important ingredient was missing in the Sommerfeld model?

(b) Outline how the Bloch theory systematically explained the missing features of part (a). Outline specifically how it explained the existence of p-type semiconductors.

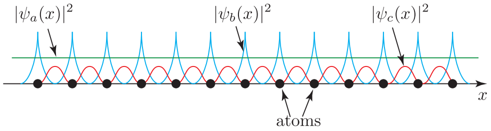


Fig. 9.17 Probability density of wavefunctions.

(c) Fig. 9.17 shows the probability density of three wavefunctions $|\psi_a(x)|^2$ and $|\psi_b(x)|^2$, and $|\psi_c(x)|^2$ in a 1D crystal of atoms. Two densities corresponds to Bloch functions of $|s\rangle$ orbital overlaps and $|p\rangle$ orbital overlaps, and the third to a free–electron wavefunction. Identify and explain which one is which.

(d) Find the total energy of all electron states in a filled 1D band of bandstructure $E(k) = E_0[1 - \cos(ka)]$, where a is the lattice constant.

(e) Find the crystal momentum $\hbar k$ and the group velocity v_g of the k-state that has the highest velocity in the band of part (d).

(f) Explain why the crystal momentum $\hbar k$ in a energy bandstructure $E(k)$ in the nearly free–electron model when the crystal potential is exactly zero is indeed the momentum of an electron in state k, but if the potential is finite, $\hbar k$ is *not* the physical momentum of state k.

(9.3) Velocity and current of Bloch states

The famous relation for the velocity of a Bloch state $\mathbf{v} = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k})$ has historically been derived and re-derived in many ways. Though some books would have one believe it is simple, it does hide several layers of complexity. And continues to spring surprises today, nearly 100 years after Bloch!

(a) Old-fashioned way (integrate and differentiate): Let us together prove the assertion that the velocity term that appears for Bloch states in the expression for the current in quantum mechanics is the *group velocity* $\mathbf{v}_g(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k})$. Re-derive each of the steps below. The definite energy wavefunctions of electrons in crystals are Bloch functions $\Psi(\mathbf{k}, \mathbf{r}, t) = \psi_E(\mathbf{k}, \mathbf{r}) e^{-iE(\mathbf{k})t/\hbar}$, where $\psi_E(\mathbf{k}, \mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u(\mathbf{k}, \mathbf{r})$ with $\mathbf{k}$ the Bloch wavevector in the reduced zone and $u(\mathbf{k}, \mathbf{r} + \mathbf{a}) = u(\mathbf{k}, \mathbf{r})$. $E(\mathbf{k})$ are the eigenvalues of the Bloch states, and constitute the electron bandstructure. To obtain the group velocity, we first derive the following useful identity for Bloch states:

$$\frac{m_e}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}) = \int d^3\mathbf{r}\Psi^\star \hat{\mathbf{p}}\Psi = -\int d^3\mathbf{r}\Psi\hat{\mathbf{p}}\Psi^\star \quad (9.69)$$

The identity is evidently dimensionally correct since both sides have units of momentum. The middle and far right are complex conjugate relations of each other, whereas the left side is real. Note that the left side involves a gradient of $E(\mathbf{k})$ in $\mathbf{k}$-space, whereas the middle and right sides involve gradients of $\Psi, \Psi^\star$ in the $\mathbf{r}$-space.
To arrive at this identity, write the Schrödinger equation for the Bloch states as

$$-\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}}\psi_E(\mathbf{k}, \mathbf{r})e^{-iE(\mathbf{k})t/\hbar} = [E(\mathbf{k}) - V]\psi_E(\mathbf{k}, \mathbf{r})e^{-iE(\mathbf{k})t/\hbar}. \quad (9.70)$$

The time evolution cancels since Bloch states are stationary states. We now take a gradient in the $\mathbf{k}$-space. Since $[\nabla_{\mathbf{r}}, \nabla_{\mathbf{k}}] = 0$, and $\nabla_k\psi_E = i\mathbf{r}\psi_E + e^{i\mathbf{k}\cdot\mathbf{r}}\nabla_{\mathbf{k}} u = i\mathbf{r}\psi_E + \mathbf{X}$ we have

$$-\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}}(i\mathbf{r}\psi_E + \mathbf{X}) = [E(\mathbf{k}) - V](i\mathbf{r}\psi_E + \mathbf{X}) + [\nabla_{\mathbf{k}} E(\mathbf{k})]\psi_E. \quad (9.71)$$

We now use the identity for Bloch eigenstates:

$$\nabla^2_{\mathbf{r}}(\mathbf{r}\psi_E) = \mathbf{r}\nabla^2_{\mathbf{r}}\psi_E + 2\nabla_{\mathbf{r}}\psi_E. \quad (9.72)$$

in Equation 9.71 and rearrange to obtain

$$\underbrace{-\frac{\hbar^2}{2m_e}(i\mathbf{r}\nabla^2_{\mathbf{r}}\psi_E)}_{1} - \frac{i\hbar^2}{m_e}\nabla_{\mathbf{r}}\psi_E = \underbrace{[E(\mathbf{k}) - V](i\mathbf{r}\psi_E)}_{1} + [E(\mathbf{k}) - (\underbrace{-\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}} + V)}_{\hat{H}}]\mathbf{X} + [\nabla_{\mathbf{k}} E(\mathbf{k})]\psi_E. \quad (9.73)$$

The terms indicated as "1" constitute the Schrödinger equation for Bloch eigenstates multiplied by $i\mathbf{r}$, and hence cancel. Multiplying by $\psi^\star_E$ on the left and integrating over all space, the second term on the right vanishes since $\psi^\star_E$ is an eigenstate. Then, we are left with

$$\nabla_{\mathbf{k}} E(\mathbf{k})\int d^3\mathbf{r}\psi^\star_E\psi_E = \nabla_{\mathbf{k}} E(\mathbf{k}) = -\frac{i\hbar^2}{m_e}\int d^3\mathbf{r}(\psi^\star_E\nabla_{\mathbf{r}}\psi_E), \quad (9.74)$$

which proves the identity in Equation 9.69 since $\hat{\mathbf{p}} = -i\hbar\nabla_{\mathbf{r}}$. The probability current density carried by a Bloch state is given by

$$\mathbf{j} = \frac{1}{2m_e}(\Psi^\star\hat{\mathbf{p}}\Psi - \Psi\hat{\mathbf{p}}\Psi^\star), \quad (9.75)$$

from which we obtain the *group velocity* of the Bloch state with help of the identity in Equation 9.69 as

$$\boxed{\mathbf{v}_g(\mathbf{k}) = \frac{1}{2m_e}\int d^3\mathbf{r}(\Psi^\star\hat{\mathbf{p}}\Psi - \Psi\hat{\mathbf{p}}\Psi^\star) = \frac{\nabla_{\mathbf{r}} E(\mathbf{k})}{\hbar}.} \quad (9.76)$$

The simplicity of the result $\mathbf{v}_g(\mathbf{k}) = \nabla_{\mathbf{r}} E(\mathbf{k})/\hbar$ is one of the central reasons why the $\mathbf{k}$-space is where one should investigate transport properties.

(b) The group velocity formula may be obtained by the application of the Hellman–Feynman theorem outlined in Chapter 7 Section 7.5. Show how the Bloch theorem and treating the wavevector $\mathbf{k}$ as a parameter with the Bloch theorem gives you the expression for the group velocity.

(9.4) Quantum transport of Bloch electrons

We derived the rather remarkable results that if we know the bandstructure $E(k)$ of any crystal, the velocity of an electron in state k is its group velocity $v_g(k) = \frac{1}{\hbar}\frac{dE(k)}{dk}$ with suitable generalization to higher dimensions, and the motion of the states in k-space is given simply by $F = \hbar\frac{dk}{dt}$ where F is the force. In this problem, consider an electron on a ring of length L moving through an extremely weak periodic crystal potential of lattice constant a such that the periodicity of the crystal is imprinted on the electron wavefunctions.

(a) Show that in the limit of a vanishing crystal potential, the wavefunction of the electron must be of the form $\psi(x) = \frac{1}{\sqrt{L}} e^{i(k+G)x}$, where $G = \frac{2\pi}{a} n$ is a reciprocal lattice vector with $n = ... -1, 0, +1, ...$

(b) The Bloch wavefunction is of the form $\psi_k(x) = e^{ikx} u(x)$ such that $u(x+a) = u(x)$. Show that the electron wavefunction of part (a) is in the Bloch form. Identify $u(x)$ and prove that indeed $u(x+a) = u(x)$.

(c) Show that the allowed electron eigenvalues are $E(k) = \frac{\hbar^2 (k+G)^2}{2m_e}$. Write expressions for several values of G, and sketch the corresponding nearly free–electron energy bands for $-3\frac{\pi}{a} \leq k \leq +3\frac{\pi}{a}$. Show on this bandstructure why at any k, there are several possible energy states with various group velocities.

(d) If an electron starts at $k = 0$ in the band indexed by $G = 0$ and experiences a force F, find the time it will take to reach an energy level degenerate with the band indexed by $G_2 = -\frac{4\pi}{a}$. Can the electron jump or transition to the other bands in the absence of a periodic potential?

(e) What does a periodic potential do to the bandstructure of part (c) and to the transport of part (d)?

(9.5) **The elusive Bloch oscillator**

In a fictitious two dimensional crystal, the bandstructure of the lowest band with a square lattice (lattice constant a) is given by

$$E(k_x, k_y) = -E_0 \cdot [\cos k_x a + \cos k_y a]. \qquad (9.77)$$

(a) Make a semi-quantitative contour plot of constant energies in the reduced Brillouin zone, and highlight energies $E = 0, \pm E_0$.

(b) Make a semi-quantitative plot of the effective mass in the (1,0) or x-direction, and the (2,1) direction in the reduced Brillouin zone.

(c) An electron is initially located at $\mathbf{k} = 0$ in $\mathbf{k}$-space, and $\mathbf{r} = 0$ is real space. At $t = 0$, a force $\mathbf{F}$ (from an electric field) is turned on which points in an oblique direction, such that $F_x = 2F_y$. Show the trajectory of the electron through the reduced zone in the $\mathbf{k}$-space, including umklapp processes, for the time interval $0 \leq t \leq 4T$, where

$$T = \frac{\pi\hbar}{aF_x}. \qquad (9.78)$$

(d) Calculate and plot the x- and y-components of the velocity and the position of the electron, all functions of time, for $0 \leq t \leq 4T$.

(e) Make a graph of the trajectory of the electron in the $x - y$ plane of real space.

(f) Explain the phenomena in words.

If you have done the problem correctly, you will realize that *very mysteriously*, the electron is oscillating in real space in the presence of an *Electric* field! A DC electric field will lead to ac oscillation power – this idea is called a Bloch oscillator. The actual implementation has remained elusive in semiconductor quantum structures. They promise very high frequency (THz) output power, only if someone can make them!

(9.6) **Momentum matrix elements of Bloch states**

In this problem, we get more familiar with Bloch states and the matrix elements between them and for momentum and space coordinates.

(a) Show that for any two Bloch states labeled by $|n\mathbf{k}\rangle = \psi_{n,\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r})$, we must have

$$\langle n\mathbf{k}|n\mathbf{k}'\rangle = \delta_{\mathbf{k},\mathbf{k}'}, \qquad (9.79)$$

which means that within the same band (same n), Bloch states with different crystal momentum $\hbar\mathbf{k}$ are orthogonal to each other.

(b) Similarly, show that

$$\langle n\mathbf{k}|m\mathbf{k}\rangle = \delta_{\mathbf{n},\mathbf{m}}, \qquad (9.80)$$

which means that Bloch eigenstates with the same $\mathbf{k}$ but belonging to different bands $n \neq m$ are orthogonal to each other.

(c) Combining the two results above, show that

$$\langle n\mathbf{k}|m\mathbf{k}'\rangle = \delta_{\mathbf{n},\mathbf{m}}\delta_{\mathbf{k},\mathbf{k}'}, \qquad (9.81)$$

whose meaning should be transparent.

(d) Show that the momentum matrix element

$$\langle n\mathbf{k}|\hat{\mathbf{p}}|m\mathbf{k}\rangle = i m_e \omega_{mn} \langle n\mathbf{k}|\mathbf{r}|m\mathbf{k}\rangle, \qquad (9.82)$$

which means that the momentum matrix element between Bloch states is related to the dipole matrix element by a factor proportional to the energy separation between the two states $\hbar\omega_{mn} = E_m - E_n$.

(9.7) **Berry phase and Berry curvature of Bloch states**
In this question, we will explore a few properties of the effect of Berry phases on the electronic states of crystals.

(a) What sort of symmetries are necessarily broken for a crystal to exhibit Berry phase in the electron eigenfunctions?

(b) Discuss what sort of crystals can exhibit non-zero Berry phases. One of the most dramatic effects of the Berry phase is the existence of spontaneous polarization in a crystal. Explore and explain what applications have resulted from exploiting the large and robust Berry phases that lead to spontaneous polarization in semiconductor crystals, and the reason for their robustness.

(c) Consider the Hall effect. Explain why the ”anomalous” velocity term in the presence of a Berry phase leads to an extra transverse Hall voltage, **even in the absence of an external magnetic field!**

Crystal Clear: Bandstructure of the Empty Lattice

10

The goals of this chapter are to understand:

- How the Bloch theorem combined with the symmetry of the periodic lattice leads to bands,
- How the crystal periodicity is experimentally measured by diffraction of light and of particles, and
- How to calculate the nearly free-electron bandstructure of *any* crystal by just knowing the periodicity and lattice constant, and with complete disregard to which atoms are in the crystal.

10.1 Diffraction as a sharp eye 205
10.2 Bragg diffraction condition 207
10.3 Broken symmetries and physical laws 208
10.4 Bravais lattices 209
10.5 Nearly free–electron bandstructure 211
10.6 Chapter summary section 215
Further reading 215
Exercises 215

10.1 Diffraction as a sharp eye

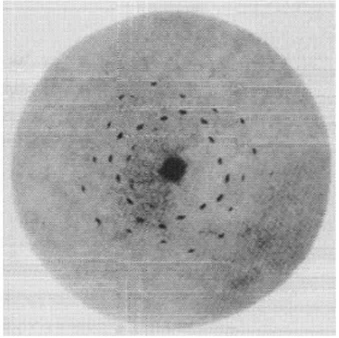


Fig. 10.1 The famous "Fig 5", showing the diffraction spots observed by Max von Laue in Munich due to Roentgen's X-rays bouncing off a crystal of ZnS, a semiconductor.

Diffraction of both waves (and particles, by the wave-particle duality) serves as a sharp eye into crystals and their periodicity. In 1912, Laue and collaborators in Munich aimed Roentgen's X-ray beams for the first time on to a crystal of the semiconductor ZnS. They observed a rather remarkable pattern on a photographic plate in the reflected beam – it is shown in Fig. 10.1. This was for the first time X-ray light was scattered off of a *crystal*, and the diffraction spots showed for the first time the evidence that the atoms inside a crystal were not arranged at random, but in a periodic pattern that could act as a grating for light waves. We have discussed earlier how electron waves diffract off the atoms in a crystal, but photons do just the same – and as in most cases, light came before electrons in revealing the nature of matter. The Bloch and Wilson theory of bands and electron waves was developed well after the breakthrough of X-ray diffraction of crystals. In fact, Laue (Fig. 10.2) and Ewald went ahead with the experiment in spite of Sommerfeld's opposition to the experiment on the grounds that the vibration of atoms means that they are never really periodic, and it should wash out any diffraction[1].

Consider the generic wave equation

$$\nabla^2\phi - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\phi = 0, \tag{10.1}$$

---

[1] Fig. 10.1 and all X-ray spectra measured since then proves the spectra are *not washed out* due to random vibrations: see Exercise 10.6 for the quantitative resolution of this conundrum!

in which v is the velocity of the wave, and ϕ is a measure of the field the wave represents. If the field is monochromatic, meaning of a single frequency ω such that $\phi = \phi_\omega(r)e^{i\omega t}$, the wave equation becomes

$$\nabla^2\phi_\omega + \frac{\omega^2}{v^2}\phi_\omega = 0 \implies \boxed{v^2\nabla^2\phi_\omega + \omega^2\phi_\omega = 0,} \qquad (10.2)$$

where we have denoted $\phi_\omega = \phi_\omega(r)$. Note that we have intentionally avoided specifying what sort of field ϕ is. That is because Equation 10.2 is a rather general wave equation that can equally well represent a beam of light, or electrons, neutrons, or what have you. For example, if $v = c$ is the speed of light and $\omega = ck$, it is Maxwell's wave equation, and ϕ_ω is the electric (or magnetic) field of an electromagnetic wave. On the other hand if we associate $v^2 = \frac{\hbar^2}{2m}$ and $\omega^2 = E$, the equation becomes $-\frac{\hbar^2}{2m}\nabla^2\phi_\omega = E\phi_\omega$. This is Schrödinger's wave equation for a quantum particle of mass m in a potential $V(r) = 0$, and the field $\phi_\omega(r) = \psi(r)$ is the *wavefunction* of the particle[2]. Thus Equation 10.2 can represent both light or electron wave propagating in vacuum with field solutions as plane waves of the form $\phi_\omega(\mathbf{r}) = A_\mathbf{k}e^{i\mathbf{k}\cdot\mathbf{r}}$.

Fig. 10.2 Max von Laue, with Peter Ewald, in 1912 discovered that X-ray beams diffract of crystals, providing for the first time a means to look into the atomic structure of matter. He was awarded the Nobel Prize for this work.

Now a crystal of semiconductor, metal, or insulator differs from vacuum by the periodic array of atoms. The atomic potential enters the wave equation as $V(\mathbf{r}) = V(\mathbf{r}+\mathbf{R})$, where $\mathbf{R}$ is the lattice constant. Because of the periodicity, we can write $V(\mathbf{r}) = \sum_\mathbf{G} \mathcal{V}_\mathbf{G} e^{i\mathbf{G}\cdot\mathbf{r}}$, where $\mathbf{G}$ is a reciprocal lattice vector. Note that $V(0) = \sum_\mathbf{G} \mathcal{V}_\mathbf{G}$. In the presence of this potential, the wave equation is

$$v^2\nabla^2\phi_\omega + \omega^2\phi_\omega = V\phi_\omega. \qquad (10.3)$$

The exact solutions to this wave equation for electrons yield the Bloch waves as discussed in Chapter 9. But if we consider the periodic potential as a *perturbation* to the free wave solutions, it will scatter an unperturbed electron plane wave in state $|\mathbf{k}\rangle = \frac{1}{\sqrt{\Omega}}e^{i\mathbf{k}\cdot\mathbf{r}}$ into a state $|\mathbf{k}'\rangle = \frac{1}{\sqrt{\Omega}}e^{i\mathbf{k}'\cdot\mathbf{r}}$ with an amplitude $\langle\mathbf{k}'|V(\mathbf{r})|\mathbf{k}\rangle$, and intensity[3]

$$\boxed{|\langle\mathbf{k}'|V(\mathbf{r})|\mathbf{k}\rangle|^2 = |\sum_\mathbf{G} \mathcal{V}_\mathbf{G}\Big(\int \frac{d^3r}{\Omega}\cdot e^{i(\mathbf{k}-\mathbf{k}'+\mathbf{G})\cdot\mathbf{r}}\Big)|^2.} \qquad (10.4)$$

The physical meaning of this scattering function is transparent: consider the sum – if $\mathbf{k}-\mathbf{k}'+\mathbf{G} \neq 0$, the complex exponential oscillates as $\mathbf{r}$ is varied over space, and the sum is zero. But if $\mathbf{k}-\mathbf{k}'+\mathbf{G} = 0$, $e^{i0} = 1$, the integral is 1, and the sum $S(\mathbf{k}' = \mathbf{k}+\mathbf{G}) = |\sum_\mathbf{G} \mathcal{V}_\mathbf{G}|^2 = V(0)^2$ is very large (the crystal potential is always large and negative, so $V(0) \neq 0$). If an electron wave moves inside the crystal, the above equation indicates that the free–electrons with wavevectors $\mathbf{k} = \pm\frac{1}{2}\mathbf{G}$ are most strongly scattered – this is exactly what we have identified earlier as the edges of the Brillouin zones: in 1D they are the points $\pm\pi/a$ in k-space, but in 2D they are lines, and in 3D the BZ edges are planes bisecting reciprocal space lattice points, the smallest of which is the first BZ that hosts the same number of **k**-points as the number of unit cells in the entire macroscopic crystal. The edges are given by

---

[2] We must be careful in limiting the analogy: The time-independent Equation 10.2 applies to waves and particles, but Equation 10.1 is not in the form of the time-dependent Schrödinger equation, which is of the form $\frac{\partial\psi}{\partial t} = D\nabla^2\psi$. This form is of a *diffusion* equation, but with a twist: it has a *complex* diffusion coefficient $D = \frac{i\hbar}{2m_e}$.

[3] In general, one writes the scattering intensity as $V(\mathbf{k},\mathbf{k}') = V(\mathbf{k}-\mathbf{k}') = V(\mathbf{q})$, and breaks it up into the scattering potential due to each identical but individual scatterer $V_0(\mathbf{q})$, with scatterers being placed in space at various locations $\mathbf{r} = \mathbf{R}_n$. Then, the scattering intensity is $V(\mathbf{q}) = V_0(\mathbf{q})S(\mathbf{q})$, where $V_0(\mathbf{q})$ is called the **form factor** which does not depend on the structure but the nature of the potential itself (say due to each atom or basis of atoms in a crystal), and $S(\mathbf{q}) = \sum_n \frac{1}{\sqrt{N}}e^{i\mathbf{q}\cdot\mathbf{R}_n}$ the **structure factor**, which does not depend on the nature of the potential, but only on how the potential is arranged in space, i.e., the structure.

$$\mathbf{k}-\mathbf{k}'+\mathbf{G}=0 \implies \mathbf{k}'=\mathbf{k}+\mathbf{G} \underbrace{\rightarrow}_{\text{BZ edges}} \mathbf{k}=-\frac{1}{2}\mathbf{G},\mathbf{k}'=\frac{1}{2}\mathbf{G}. \tag{10.5}$$

The above scattering is responsible for the opening of bandgaps in the energy spectrum at these edges of the BZ, when the free–electron wavelength becomes commensurate with the lattice constants $\frac{2\pi}{\lambda} = n\frac{2\pi}{2a} \rightarrow n\lambda = 2a$.

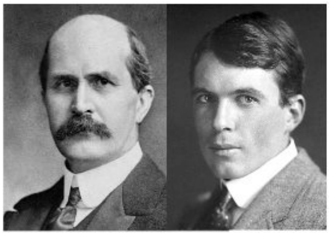


Fig. 10.3 Father and son Braggs. They were awarded the 1915 Nobel prize for the X-ray study of crystals. William Bragg also realized seeing Laue's (Fig. 10.1) X-ray photograph that the crystal structure of ZnS was not simple cubic, but face-centered cubic (FCC)!

10.2 Bragg diffraction condition

As we have stressed, Equation 10.2 is true both for electrons and photons. If instead of electrons, X-ray photons are incident on the crystal, they will suffer exactly the same fate of diffraction. This is because the electron clouds around the nuclei of the atoms of the crystal create a periodically changing electronic charge density of the form $n(\mathbf{r}+\mathbf{R}) = n(\mathbf{r})$, where $\mathbf{R}$ is a real-space lattice vector. In the microscopic picture, electrons of the atoms in the crystal oscillate in phase with the electric field component of the X-ray electromagnetic wave, and re-radiate the wave as a dipole – this is called **Thomson scattering**. Since the strength of such scattering is inversely proportional to the mass of the charged particle, the light electron clouds dominate, and the far too heavy nuclei contribute very little to the re-radiation. Fig. 10.4 indicates the experiment: the X-ray photons are incident on a crystal with wavevector **k**, and reflect with wavevector **k**′ to a screen (or photographic plate), on which it produces a scattering pattern: this pattern is precisely given by the boxed Equation 10.4. The condition for strong peaks in the case of **elastic scattering** of photons from atomic planes distance d apart is found using $\mathbf{k}'-\mathbf{k}=\mathbf{G}$:

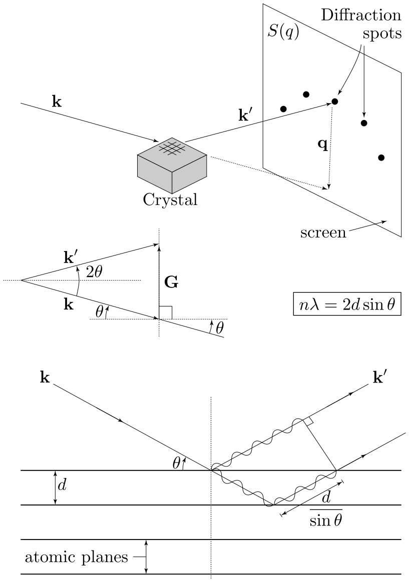


Fig. 10.4 Bragg law for diffraction of X-rays from a crystal.

$$\mathbf{k}\cdot(\mathbf{k}'-\mathbf{k})=\mathbf{k}\cdot\mathbf{G} \implies \mathbf{k}\cdot\mathbf{k}'-k^2=\mathbf{k}\cdot\mathbf{G}$$

$$\underbrace{\implies}_{\text{elastic},|\mathbf{k}|=|\mathbf{k}'|=k} \quad k^2\cos(2\theta)-k^2=k(n\frac{2\pi}{d})\underbrace{\cos\left(\frac{\pi}{2}+\theta\right)}_{-\sin\theta}$$

$$k=\frac{2\pi}{\lambda} \implies \frac{2\pi}{\lambda}\underbrace{(1-\cos(2\theta))}_{2\sin^2(\theta)}=n\frac{2\pi}{d}\sin\theta \implies$$

$$\boxed{n\lambda = 2d\sin\theta.} \tag{10.6}$$

This relation was first identified by Lawrence Bragg (Fig. 10.3), shown with his father William Bragg, who constructed the X-ray diffractometer. Fig. 10.4 shows the diffraction condition pictorially. Because the X-ray photons are not sensitive to the same crystal potential $V(\mathbf{r})$ that scatter electrons, but to the electron density $n(\mathbf{r})$, the X-ray scattering intensity is given by an almost identical expression as Equation

10.4:

$$\boxed{S(\mathbf{q}) = |\underbrace{\int d^3r \cdot n(\mathbf{r}) \cdot e^{i(\mathbf{q}\cdot\mathbf{r})}}_{F(\mathbf{q})}|^2 = |F(\mathbf{q})|^2,} \tag{10.7}$$

and is commonly called the **structure factor**. The scattering amplitude $F(\mathbf{q})$ identified as the Fourier transform of the electron density, is called the **form factor**. The diffraction spots due to elastic scattering form on the screen at $S(\mathbf{q})$ where $\mathbf{q} = \mathbf{k} - \mathbf{k}' = \pm\mathbf{G}$ as indicated in Fig. 10.4. The Bragg condition may also be pictorially derived by realizing that there is constructive interference when the path length difference between two wavefronts is an integer number of wavelengths, or $n\lambda = \frac{2d}{\sin\theta} - \frac{2d}{\tan\theta}\cos\theta \implies n\lambda = 2d\sin\theta$. Since $|\sin\theta| \leq 1$, the X-ray wavelength should be of the order, or less than, the lattice constant for the method to work. Historically, the **Laue method** involves range of incident X-ray wavelengths, the **Bragg method** involves a monochromatic incident X-ray beam but changes θ by turning the crystal, and the **Debye–Scherrer method** is performed on many small randomly oriented crystalline particles, commonly referred to as powder diffraction (e.g. see Exercise 10.7). All the three methods find heavy use today to characterize the crystal structure of semiconductor nanostructures used in electronic and photonic devices. In fact the power of X-ray crystallography has been a sharp eye into the structure of matter in a diverse range of fields: in analyzing the structure of polymers and soft matter, and was instrumental in decoding the blueprint of life itself, the DNA (see Exercise 10.9).

Fig. 10.5 Woldemar Voigt realized the connection between symmetry and the physical properties of crystals. He introduced much of the notations used in crystallography today.

Fig. 10.6 Emmy Noether was one of the most influential mathematicians of the recent past. She proved the mathematical theorems that relate symmetries to conservation laws in physics.

10.3 Broken symmetries and physical laws

Before X-ray diffraction became available in the early 20th century, generations of crystallographers had tried to find a reason for the shapes of naturally occurring or chemically synthesized crystals. In the late 1800s, Ernst Neumann and in particular his student and successor, Woldemar Voigt (fig. 10.5)connected the symmetries in crystal structure to their properties (the so-called structure-property relationship). Voigt's notations are still in use today- he introduced them in his 1910 book, considered to be a landmark in the history of science because it also introduced the philosophy of analysis through symmetry. Pierre Curie realized that breaking of symmetries in crystals gives rise to physical properties such as piezoelectricity or pyroelectricity. This rather powerful and far reaching idea – that **broken symmetries are the origin of physical laws** – permeates all of modern physics (see Fig. 10.6). It is humbling to realize that it was discovered not by abstract mathematics, but in the effort to understand the physical properties of crystals measured in laboratories. In other words, a crystal represents the physical embodiment of symmetries – that you can hold in your hand, and *measure* broken symmetries in them to your heart's content!

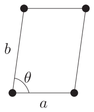
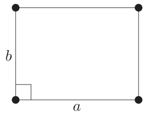
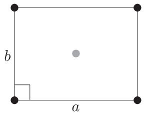
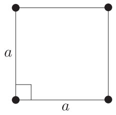
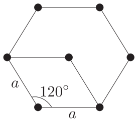
	Primitive	Centered	Edge lengths	Axial angles	Area
Monoclinic			$a \neq b$	$\theta \neq 90°$	$ab \sin\theta$
Orthorhombic			$a \neq b$	$\theta = 90°$	ab
Tetragonal			$a = b$	$\theta = 90°$	a^2
Hexagonal			$a = b$	$\theta = 120°$	$\frac{\sqrt{3}}{2}a^2$

Fig. 10.7 The five 2D Bravais Lattices.

10.4 Bravais lattices

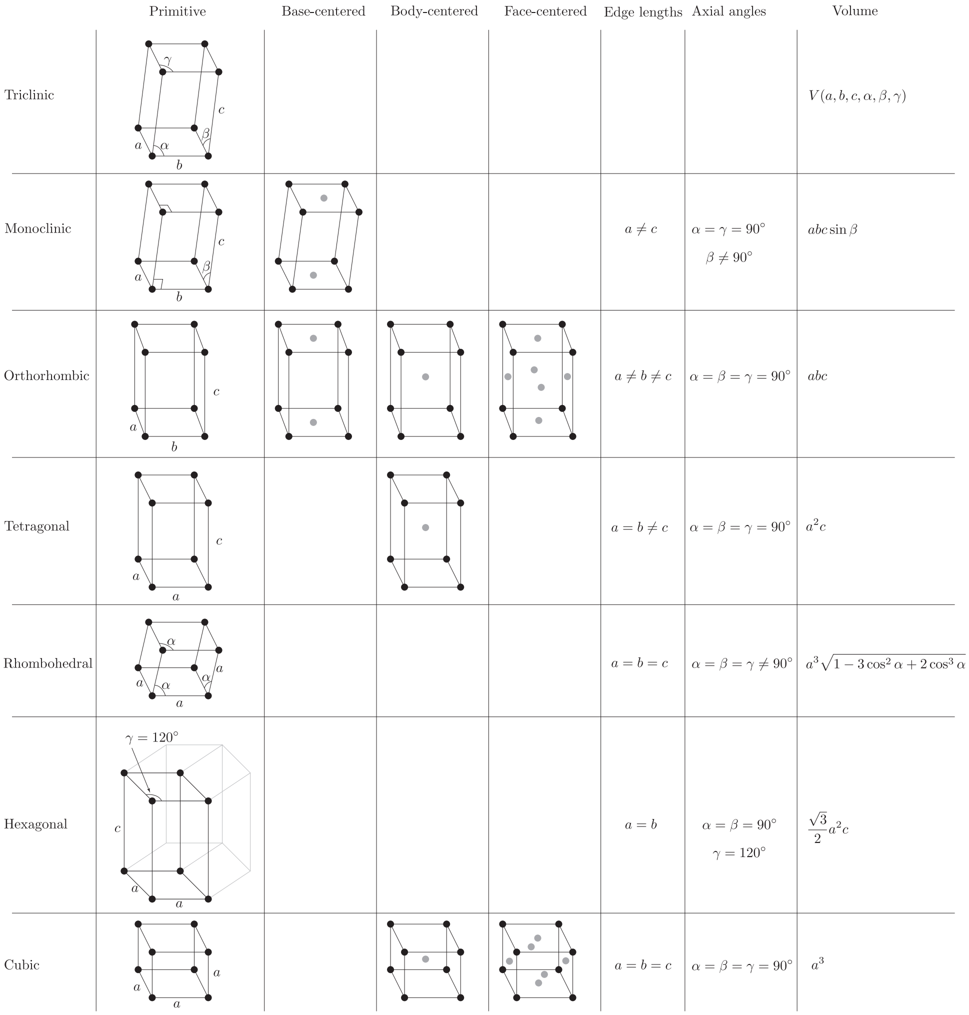
What are the fundamental repeating units, or Wigner–Seitz cells that can be stacked by translating in space to fill the space completely? One may expect that an infinite number of possible primitive surfaces (for 2D) or volumes (for 3D) exist that can be repeated and stacked to fill all space. However, this is not correct. The French scientist Auguste Bravais (Fig. 10.8) proved that to fill all space, nature restricts us to **only 5 choices of Wigner–Seitz cells in 2D**, and to **only 14 choices of Wigner–Seitz cells in 3D!** In other words, the lattice of every 3D crystal must take one of these 14 choices[4]. These are shown in Figures 10.7 and 10.9.

Fig. 10.8 French scientist Auguste Bravais identified 14 fundamental repeating units for any 3D crystal, which are named the **Bravais lattices** in his honor. Note that this was discovered *before* X-ray diffraction measurements confirmed it in real crystals!

In 2D crystals such as graphene, BN, or MoS_2, the hexagonal lattice is common. Among these, graphene and BN monolayers each have two-atom basis: for graphene both atoms are C, whereas the basis of monolayer BN has one B and one N atom. For a transition metal dichalcogenide 2D crystal such as MoS_2, to each lattice point is attached a basis of a S-Mo-S atomic triplet. If several 2D layers are stacked on top of each other with a periodic relationship between atoms of various layers, their lattice evolves to 3D, and must be one of the 14 3D Bravais lattices. For example, for graphite and hBN, the 3D lattice is hexagonal close packed (HCP), which is shown as hexagonal primitive in Fig. 10.9.

Among the 3D crystals, the group-IV elemental semiconductors, diamond, silicon and germanium take the Face-Centered Cubic (FCC) Bravais lattice structure. A large number of cubic III-V semiconduc-

[4] If rotations are also allowed in addition to translation, the number of lattices extends to 230 space groups! For example, an interesting 3D lattice is formed by stacking 2D lattices rotated by $\theta_n = 2\pi/n$ where n is an integer. The few-layer units of these crystals exhibit Moiré patterns in their structure and exotic electronic phenomena.



	Edge lengths	Axial angles	Volume
Triclinic			$V(a,b,c,\alpha,\beta,\gamma)$
Monoclinic	$a \neq c$	$\alpha = \gamma = 90°$ $\beta \neq 90°$	$abc \sin\beta$
Orthorhombic	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90°$	abc
Tetragonal	$a = b \neq c$	$\alpha = \beta = \gamma = 90°$	$a^2 c$
Rhombohedral	$a = b = c$	$\alpha = \beta = \gamma \neq 90°$	$a^3\sqrt{1 - 3\cos^2\alpha + 2\cos^3\alpha}$
Hexagonal	$a = b$	$\alpha = \beta = 90°$ $\gamma = 120°$	$\frac{\sqrt{3}}{2}a^2 c$
Cubic	$a = b = c$	$\alpha = \beta = \gamma = 90°$	a^3

Fig. 10.9 The 14 3D Bravais Lattices. The general volume is $V(a,b,c,\alpha,\beta,\gamma) = abc\sqrt{1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}$.

tors such as GaAs, InP, InAs, Insb, ZnS, etc. also take the FCC Bravais lattice. As was shown in Fig. 9.7, the basis of these crystals consists of two identical atoms, such as Si-Si at $(0,0,0)$ and $a(\frac{1}{4},\frac{1}{4},\frac{1}{4})$, where a is the side of the cubic cell, 5.43 Å for silicon. The compound semiconductors with the FCC lattice structure sometimes go under the name zinc blende, after ZnS which was the first crystal on which X-ray diffraction was performed, and the crystal structure was first identified (see Fig. 10.1). The identical atoms in the basis of the elemental semiconductors in the cubic Bravais lattice have the same ionicity, and hence these semiconductors are not piezoelectric. However, since the basis atoms of the zinc-blende semiconductors have a group III and a group V atom (or group II and group VI atoms), the difference in electronegativity leads to an unequally shared charge cloud between them. Hence, when these crystals are strained, equal and opposite charge planes develop in the rows of slightly positively and slightly negatively charged atoms, leading to **piezoelectricity**[5].

The nitride semiconductors GaN, InN, and AlN take the wurtzite crystal structure under ambient conditions. The underlying Bravais lattice is the hexagonal close packed (HCP) as seen in Fig. 10.9. There are four atoms in the basis: for example, two Ga and two N atoms for GaN. The lattice constants are $a = 3.19$ Å and $c = 5.189$ Å for GaN, leading to a ratio $\frac{a}{c} = 0.614 = \sqrt{0.377}$. The ideal HCP lattice has a close-packing ratio of $\frac{a}{c} = \sqrt{\frac{3}{8}} = \sqrt{0.375}$. The deviations of the equilibrium crystal structures compared to the ideal HCP lattice has strong effects on the electronic polarization: spontaneous and piezoelectric polarization in wurtzite crystals such as the nitride semiconductors. In particular, they are **strongly piezoelectric** (roughly 10 times stronger than GaAs and other zinc blende semiconductors), and also have **spontaneous polarization**, a property that is missing in zinc blende crystals due to their cubic lattice. In other words, the III-nitride crystals have broken inversion symmetry along the c-axis, and thus have either metal (Ga, Al, In)-polar surface or N-polar surface at one end, and the other polarity at the other end. Because of spontaneous polarization, these semiconductor crystals also exhibit internal electric fields, similar to ferroelectrics. They are different from ferroelectrics since in the wurtzite nitride crystals the polarization orientation is locked with the crystal orientation and cannot be flipped by external electric fields[6]. ZnO also takes the wurtzite crystal structure and has lattice constants close to GaN.

10.5 Nearly free–electron bandstructure

Since the lattice structures of the cubic and hexagonal close packed crystals have now been determined, we are in a position to completely solve for the nearly free–electron bandstructure as outlined in Equation 9.9. Let us first consider the FCC lattice. The real space lattice vectors are $\mathbf{a}_1 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}}) = a(\frac{1}{2}, \frac{1}{2}, 0)$, $\mathbf{a}_2 = a(0, \frac{1}{2}, \frac{1}{2})$, and $\mathbf{a}_3 = a(\frac{1}{2}, 0, \frac{1}{2})$ as indicated in Table 10.1. The volume of the Wigner-Seitz cell is $\Omega_r = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] = \frac{a^3}{4}$. The corresponding reciprocal space lattice vectors are $\mathbf{b}_1 = \frac{2\pi}{a}(\hat{\mathbf{k}}_x + \hat{\mathbf{k}}_y - \hat{\mathbf{k}}_z)$, $\mathbf{b}_2 = \frac{2\pi}{a}(\hat{\mathbf{k}}_x - \hat{\mathbf{k}}_y + \hat{\mathbf{k}}_z)$, and $\mathbf{b}_3 = \frac{2\pi}{a}(-\hat{\mathbf{k}}_x + \hat{\mathbf{k}}_y + \hat{\mathbf{k}}_z)$. The volume of the reciprocal unit cell is $(2\pi)^3/\Omega_r$. The reciprocal lattice vectors $\mathbf{G} = l\mathbf{b}_1 + m\mathbf{b}_2 + n\mathbf{b}_3$ are then

$$\mathbf{G} = \frac{2\pi}{a}[(l + m - n)\hat{\mathbf{x}} + (l - m + n)\hat{\mathbf{y}} + (-l + m + n)\hat{\mathbf{z}}]. \tag{10.8}$$

The nearly free–electron eigenvalues for the FCC lattice are then

$$E(k_x, k_y, k_z) = \frac{\hbar^2}{2m_e}[(k_x - G_x)^2 + (k_y - G_y)^2 + (k_z - G_z)^2]. \tag{10.9}$$

---

[5] The piezoelectric effect leads to surface acoustic waves (SAW) that can be excited by electric fields applied across electrodes on the surface of these semiconductors.

[6] These crystals are also called **pyroelectric** because the charged surfaces attract oppositely charged ions from the atmosphere and neutralize. When the crystals are heated, these adsorbed ions are desorbed, and the charges on the surface of the crystals produce an electric field: the heat produces electricity, hence pyroelectric. Long thin needle or nanowire shaped crystals oriented along the c-axis are also called **electrets** in analogy to magnets.

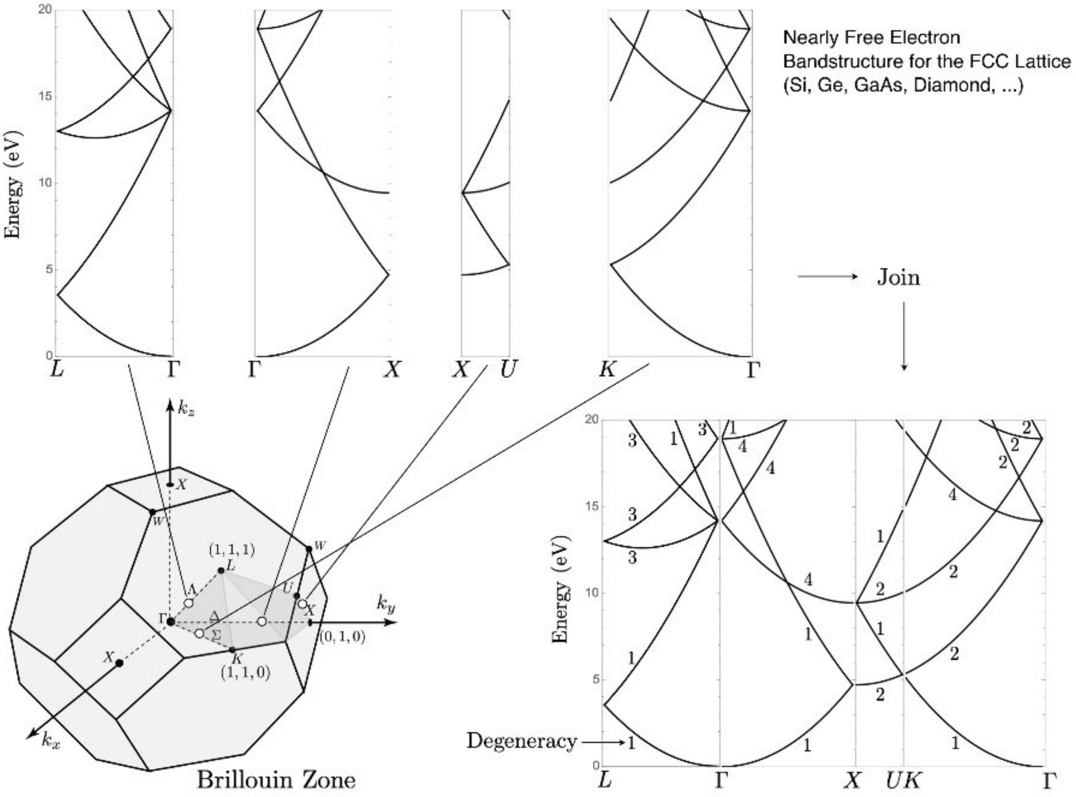


Fig. 10.10 Nearly free–electron bandstructure for the FCC lattice. Silicon, germanium, GaAs, diamond, and many semiconductors have a FCC real-space lattice with 2-atom bases. The nearly free–electron bandstructure shown here is representative of all semiconductors that share the same real-space lattice. The wedge shown in darker shade of gray is the smallest region of the BZ that contains all eigenvalues, since the BZ can be constructed by transformations of this wedge around the origin. The **k**-space path $L \to \Gamma \to X \to U \to K \to \Gamma$ is one of many possible, and is standard in the literature. Note that U and K points are identical by symmetry.

Table 10.1 FCC crystal lattice properties.

$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
$a(\frac{1}{2},\frac{1}{2},0)$	$a(\frac{1}{2},0,\frac{1}{2})$	$a(0,\frac{1}{2},\frac{1}{2})$
$\mathbf{R} = n_1\mathbf{a}_1$	$+n_2\mathbf{a}_2$	$+n_3\mathbf{a}_3$
$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
$\frac{2\pi}{a}(1,1,-1)$	$\frac{2\pi}{a}(1,-1,1)$	$\frac{2\pi}{a}(-1,1,1)$
$\mathbf{G} = l\mathbf{b}_1$	$+m\mathbf{b}_2$	$+n\mathbf{b}_3$

By choosing a set of **G** vectors with various (l, m, n), we plot the eigenvalues in Fig. 10.10. Also shown in this figure is the 3D Brillouin zone of the FCC lattice created by taking the perpendicular bisector planes of the lines connecting the origin $\Gamma = (k_x, k_y, k_z) = (0, 0, 0)$, which is a reciprocal lattice point with $(l, m, n) = (0, 0, 0)$ and all nearest and next-nearest neighbor points. The points of high symmetry are $\Gamma = (0, 0, 0)$, $X = \frac{2\pi}{a}(1, 0, 0)$, and all its permutations $\frac{2\pi}{a}(-1, 0, 0)$, $\frac{2\pi}{a}(0, \pm1, 0)$ and $\frac{2\pi}{a}(0, 0, \pm1)$, $L = \frac{2\pi}{a}(1, 1, 1)$ and all its permutations, and $K = \frac{2\pi}{a}(1, 1, 0)$ and all its permutations. The reciprocal lattice points form a BCC lattice. The BZ surfaces host six square faces perpendicular to the $\Gamma - X$ directions, and eight hexagonal faces perpendicular to the $\Gamma - L$ directions. Each **k**-point inside this BZ has a discrete set of allowed energies given by $\frac{\hbar^2(\mathbf{k}+\mathbf{G})^2}{2m_e}$ for various **G**. It is not possible to plot this 4D object in the 2D plane. It is therefore the custom to plot the eigenvalues along certain chosen 1D lines of high symmetry

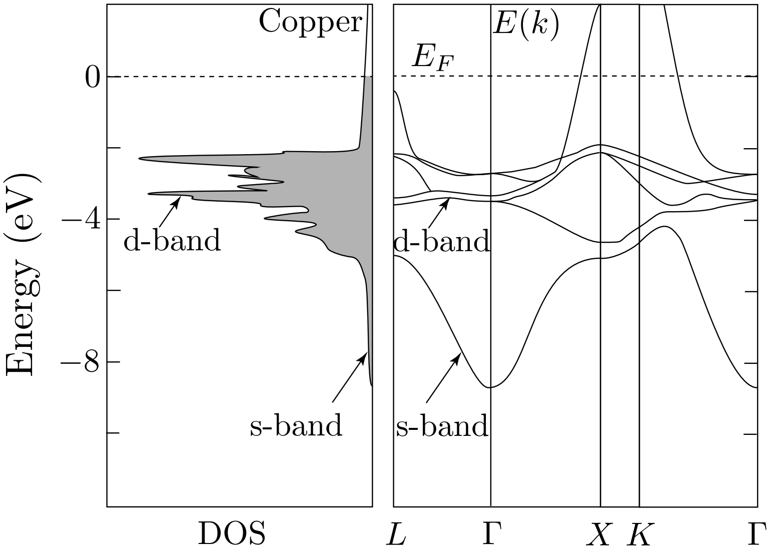


Fig. 10.11 The actual bandstructure and the density of states of FCC copper.

in the **k**-space to capture as much information about the eigenvalues as possible. Fig. 10.10 shows the bands along one such chose 1D path, along with the degeneracies of each nearly free-electron band. After plotting the bands along each section, they are joined together, though one must remember that the bands appear asymmetric around the Γ point, but they are not. This is because the $L - \Gamma$ direction is not the negative of the $\Gamma - X$ direction. Note that the two lowest eigenvalues at the Γ point are separated by ~ 15 eV [7].

Consider the actual bandstructure not of a semiconductor, but of copper, the metallic crystal shown in Fig. 10.11. Copper takes a FCC lattice, and therefore its nearly free–electron bandstructure is exactly the same in shape as shown in Fig. 10.10. First note that the real energy bands of Fig. 10.11 are plotted along the same **k**-space path as Fig. 10.10. Also note that the lowest s-band is nearly parabolic and continues at high energies, but is crossed by d-bands in-between. The s-bands are close to the nearly free-electron bandstructure model, and produce a light effective mass and the small density of states. The d-bands are much more atomic-like: they are almost flat in the $E(k)$ diagram with small dispersion and a heavy effective mass. Therefore, they produce sharply peaked density of states. When electrons are filled, the Fermi level crosses the s-bands, giving copper its characteristic high electronic conductivity.

We next consider the bandstructure of the nitride semiconductors GaN, AlN, and InN, which are used for visible and ultraviolet photonics, as well as for high-speed and high-voltage electronics. The real space lattice vectors are shown in Table 10.2. The volume of the unit cell is $\Omega_r = \frac{\sqrt{3}}{2}a^2c$. The corresponding reciprocal space lattice vectors are shown in the same table, which makes the Brillouin zone volume $(2\pi)^3/\Omega_r$. Defining $c = a/\sqrt{u}$, the dimensionless ratio $u = (\frac{a}{c})^2$ is

---

[7] When valence electrons in silicon or GaAs will fill these states, there are eight electrons/unit cell, which can completely fill four bands. The four bands will be the lowest band starting from Γ, and three at the Γ-point at ~ 15 eV, but with energy eigenvalues perturbed from the values shown. The three bands curving down from the Γ point starting around ~ 15 eV will become the valence bands, and the width of the valence band and the bottom band will remain close to ~ 15 eV. We will encounter these details in the following chapters on bandstructure.

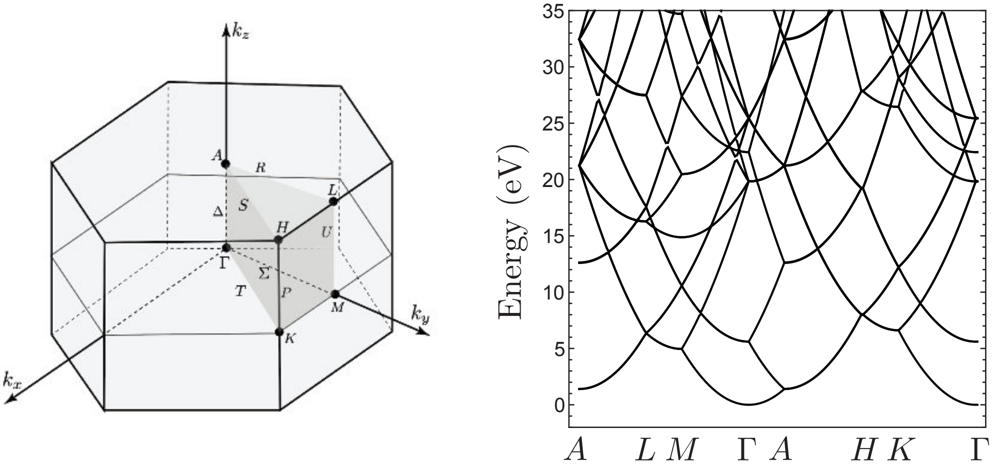


Fig. 10.12 Nearly free–electron bandstructure for the HCP lattice. GaN, AlN, InN, ZnO, ZnS, and many semiconductors have a HCP real-space lattice with four-atom bases. The nearly free–electron bandstructure shown here is representative of all semiconductors that share the same real-space lattice. The wedge shown in darker shade of gray is the smallest region of the BZ that contains all eigenvalues. The **k**-space path $A \to L \to M \to \Gamma \to A \to H \to K \to \Gamma$ is standard in the literature for HCP lattices, especially for semiconductors that have the wurtzite crystal structure.

Table 10.2 HCP crystal lattice properties.

$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
$a\hat{\mathbf{x}}$	$a(\frac{1}{2}\hat{\mathbf{x}}+\frac{\sqrt{3}}{2}\hat{\mathbf{y}})$	$c\hat{\mathbf{z}}$
$\mathbf{R}=n_1\mathbf{a}_1$	$+n_2\mathbf{a}_2$	$+n_3\mathbf{a}_3$
$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
$\frac{2\pi}{a}(\hat{\mathbf{x}}-\frac{1}{\sqrt{3}}\hat{\mathbf{y}})$	$\frac{2\pi}{a}\frac{2}{\sqrt{3}}\hat{\mathbf{y}}$	$\frac{2\pi}{c}\hat{\mathbf{z}}$
$\mathbf{G}=l\mathbf{b}_1$	$+m\mathbf{b}_2$	$+n\mathbf{b}_3$

used to characterize the semiconductor. For an ideal HCP structure, $\frac{c}{a}=\sqrt{\frac{8}{3}}$, or $u=\frac{3}{8}$. The equilibrium lattice constants of GaN, AlN, and InN differ from this ideal value, therefore it is important to keep track of the value of u in the calculations. The reciprocal lattice vectors are then given by $\mathbf{G}=l\mathbf{b}_1+m\mathbf{b}_2+n\mathbf{b}_3$, which is written in cartesian coordinates as

$$\mathbf{G}=\frac{2\pi}{a}[l\hat{\mathbf{x}}+(-\frac{l}{\sqrt{3}}+\frac{2m}{\sqrt{3}})\hat{\mathbf{y}}+(\sqrt{u}n)\hat{\mathbf{z}}]. \tag{10.10}$$

The nearly free–electron bandstructure for various values of (l,m,n) using the reciprocal lattice vectors above is shown in Fig. 10.12. Since $c>a$ the $\Gamma-A$ distance is less than $\Gamma-M$ distance. There is a lot more structure to the bands for the HCP than the FCC lattice. There are four atoms in the basis of the wurtzite crystals GaN, AlN, InN, meaning there will be sixteen valence electrons that can fill eight bands from the bottom. Note that the lowest energy eigenvalues at the Γ-point are at 0, $\sim$ 6, and 20 eV. The valence bands of the wurtzite nitrides will be found to have more structure than the FCC lattice crystals, but the total width will still be around 20 eV.

In the next three chapters, as we compute the energy bandstructures of various semiconductors, it will be useful to note that far from the high symmetry points and the BZ edges, the energy bandstructures are retain the same shapes as the nearly free–electron bands of Figs. 10.10 and 10.12. This is because electrons at those **k** only weakly experience the periodic crystal potential.

10.6 Chapter summary section

In this chapter, we have learned:

- How diffraction of photons, electrons, or neutrons from the periodic potential of a crystal enables the direction measurement of the symmetries and the periodicity of the crystal.
- The connection between broken symmetry and physical laws.
- The fact that a finite number of repeating Bravais lattices fill all space in 2D and 3D space.
- The reciprocal lattice space and Brillouin zones of a crystal are therefore determined from the lattice without regard to the specific atoms.
- How to draw the nearly free–electron bandstructure of crystals in which the electron wavefunction is still that of a free–electron, but with the requirement that it must satisfy Bloch's theorem. The energy then takes the form $E(k) = \frac{\hbar^2|\mathbf{k}\pm\mathbf{G}|^2}{2m_e}$ for all reciprocal lattice vectors $\mathbf{G}$, and therefore there are several bands and crossings.
- How to calculate the nearly free–electron bandstructures of FCC and HCP crystal structures.

Further reading

The books referred to in Chapters 8 and 9, primarily **Solid State Physics** by Ashcroft and Mermin and **Introduction to Solid State Physics** by Charles Kittel has a good introductory treatment of the nearly free-electron bandstructure. **Quantum Mechanics** by Herbert Kroemer has an illuminating discussion of the nearly free–electron model and the consequences of Bloch theorem. **Physical Properties of Crystals** by J. F. Nye is a classic tensor based introduction to the connection of physical properties with the structure and geometry of atoms.

Exercises

(10.1) **Wigner–Seitz cells and the reciprocal lattice**

Fig. 10.13 (adapted from Kroemer's text on Quantum Mechanics) shows the arrangement of atoms of a fictitious two-dimensional crystal. All circles (filled and empty) represent the same atom.

(a) Using the filled "atom" shown in black as the origin of the lattice, indicate the primitive translation vectors of this lattice. (There are several possible choices; use the simplest one, and explain.)

How many atoms are there per lattice point?

(b) Determine the reciprocal lattice vectors corresponding to this lattice, and sketch the reciprocal lattice as accurately as you can. Use the grid lines shown as an unit of measure.

(c) Construct and show the Wigner–Seitz cell, or equivalently, the Brillouin zone clearly in the reciprocal lattice plot.

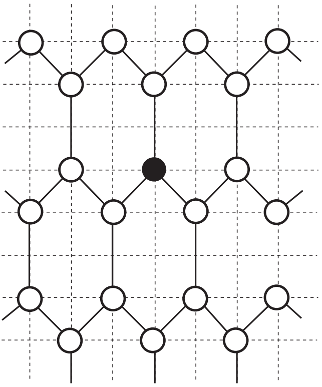


Fig. 10.13 Atomic arrangement of a fictitious 2D crystal.

(10.2) **Crystal math**

(a) Show that if $s = 0, 1, \ldots, N-2, N-1$, and a is a lattice constant, and $S_q = \sum_{s=0}^{N-1} e^{i(qa)s}$, then

$$|S_q|^2 = \frac{\sin^2(Nqa/2)}{\sin^2(qa/2)}. \qquad (10.11)$$

(b) Plot $|S_q|^2$ for $N = 30$ against q, and identify the values of q where the peaks occur.

(c) Show that the value of intensity peaks are N^2, and (Hard!) the area of each period of $|S_q|^2$ is $A = 2\pi N/a$. You might have to evaluate a complex integral.

(d) (Hard!) Show that the full width at half max (FWHM) of the peaks of $|S_q|^2$ is given by $2\Delta q \approx 5.42/Na$ (assume that Δq is small).

(e) What happens to the peaks (magnitude, area, and FWHM) as $N \to \infty$? What mathematical function best describes $|S_q|^2$ in this limit? Show by qualitative arguments that in this limit, you can derive a *very* useful result:

$$\sum_{s=0}^{N-1} e^{iqsa} = \frac{2\pi}{a} \sum_{m=-\infty}^{+\infty} \delta(q - \frac{2\pi}{a}m) \qquad (10.12)$$

Note that this result is also found in books of signal processing, Fourier transforms, etc. The chain of delta functions on the right, looks like a "bed of nails" – we will refer to this function by this name for later work.

(f) Generalize this sum to the three dimensional case; it is again a very useful result that relates the real space lattice vectors $\vec{R}_l$ to reciprocal space lattice vectors $\vec{G}$:

$$\sum_l e^{i\vec{q}\cdot\vec{R}} = \frac{(2\pi)^3}{a^3} \sum_{\vec{G}} \delta(\vec{q} - \vec{G}). \qquad (10.13)$$

(10.3) **1D crystal**

You will now make sense out of the math of Exercise 10.2. Assume a hypothetical "crystal", comprised of *identical* atoms periodically arranged in a line, with spacing a. There are N atoms in the line, making the length of the crystal $L = Na$. Now a wave (say X-ray, or high-energy neutrons) is incident on the crystal. The crystal will diffract this wave; the *amplitude* of the diffracted wave (from electromagnetic theory) is proportional to

$$S_q \propto \sum_s f_s e^{i(qa)s}, \qquad (10.14)$$

where the quantity f_s is called an "atomic form factor". It is a measure of the scattering power of the atom, i.e., it is an atomic property. Hence, it is the same for all atoms of the same kind. From the previous Exercise, show the following:

(a) **Bragg/Laue condition**: For a finite crystal, The diffracted beam peaks are along directions which meet the condition $qa = 2\pi m$.

(b) **Crystal extent**: The "larger" the size of the crystal is, the sharper and stronger are the diffraction peaks.

(c) **Crystal with basis**: If the crystal has two distinct atoms in the basis with form factors f_A and f_B, the diffraction pattern changes; some peaks disappear, and new ones appear. This is how crystal structure is revealed in diffraction measurements.

(10.4) **3D crystal**

Consider a finite-size, simple cubic crystal of $N_x \times N_y \times N_z$ atoms, with lattice spacing a.

(a) For $k_x = k_y = 0$, plot the k_z-dependence of the intensity of the scattered X-rays around the (0,0,1) point in reciprocal space. Show that there is a peak at (0,0,1) and that the peak width is finite. How is the width related to the size of the crystal N_z? Assume $N_z \gg 1$.

(b) What happens to the line width if we have a collection of crystallites, which are all oriented parallel to each other, but their positions are random?

(10.5) **Broken symmetries, scattering, structure factors**
(a) Discuss the origin of "Goldstone modes" due to broken symmetries. Give examples of Goldstone modes in semiconductor crystals.

(b) The structure factor $\mathcal{S}(\mathbf{q}) = \frac{1}{N}|\sum_j e^{i\mathbf{q}\cdot\mathbf{R}_j}|^2$ was discussed as the mathematical function that captured the degree of order (or disorder) in a crystal of lattice sites $\mathbf{R}_j$ when probed in a scattering experiment, in which the wavevector of the probe beam changes by $\mathbf{q} = \mathbf{k}_{out} - \mathbf{k}_{in}$. Show that for a 1D crystal of lattice constant a and N sites the structure factor is $\mathcal{S}(q) = \frac{1}{N}\frac{\sin^2(\frac{Naq}{2})}{\sin^2(\frac{aq}{2})}$.

(c) Plot the structure factor for a small linear crystal with $N = 10$ for $0 \leq qa \leq 10\pi$. How many satellite peaks are there between the maxima peaks? Re-plot the structure factor for $N = 100$ for $0 \leq qa \leq 10\pi$ What is the maximum peak strength? Prove that in general both the maximum peak and the number of satellite peaks between the maximum peaks $= N$.

(d) Re-Plot the structure factor for $N = 10$ for $0 \leq qa \leq 10\pi$, but now remove three randomly selected sites to represent vacancies. What is the maximum peak strength now? How many peaks are there between the maxima peaks? Repeat this exercise for $N = 100$ to find the peak height and width. Projecting from this simple exercise, would you recommend diffraction experiments to probe point defects and vacancies in semiconductors? Why or why not?

(10.6) **Order hidden inside disorder: the Debye–Waller factor**
In 1912 Max von Laue conjectured that scattering of X-rays from a crystal should result in a diffraction pattern. When he discussed this with his colleagues Sommerfeld, Wien, and others (famously in a "coffee-house" in Munich), Sommerfeld argued that any possible diffraction spots should be completely washed out because of the random thermal vibrations of the atoms from their equilibrium perfect periodic positions. However, this does not happen – the spots do appear, as was shown soon thereafter experimentally, in spite of Sommerfeld not allowing resources for the experiment! Peter Debye proved why this is so, and you will do so in this Exercise.

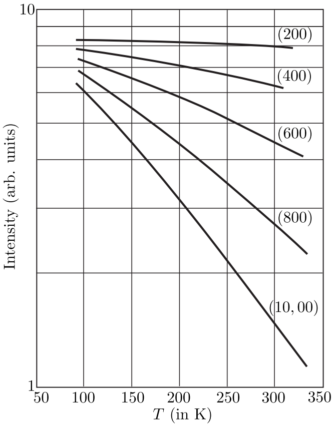


Fig. 10.14 X-ray diffraction peak dependence on temperature for an aluminum crystal (from Kittel's *Solid State Physics*).

(a) Show that if because of random thermal $T \neq 0$ vibrations the location of the atoms at lattice sites j, k at time t change to $\mathbf{R}_j \to \mathbf{R}_j + \mathbf{u}_j(t)$ and $\mathbf{R}_k \to \mathbf{R}_k + \mathbf{u}_k(t)$, the resulting structure factor because of a wave scattering off of the crystal changes to $\mathcal{S}(\mathbf{q}, T) = \mathcal{S}(\mathbf{q}, T = 0) \times \exp[i\mathbf{q} \cdot \sum_{jk}(\mathbf{u}_j(t) - \mathbf{u}_k(t))]$. Neglect the quantum zero-point vibrations in this Exercise. Meaning, at $T = 0$ K, the atoms sit still and do not vibrate.

(b) Assume a 3D crystal with atoms of mass M and their characteristics vibration frequency ω_0. Expand the finite-temperature exponential term to show that the structure factor at finite temperatures is $\mathcal{S}(\mathbf{q}, T) = \mathcal{S}(\mathbf{q}, T = 0) \times \exp[-\frac{k_B T q^2}{M\omega_0^2}]$. The exponential factor is called the **Debye–Waller factor**.

(c) Re-do the structure factor calculation in Exercise 10.5 to derive an analytical expression for the dynamical structure factor. How does the

Debye–Waller factor affect the temperature dependence of higher-order satellite peaks? Does it make intuitive sense?

(d) Fig. 10.14 shows the measured reduction in X-ray peak intensity for an aluminum crystal with temperature. Explain the experimental data *quantitatively* by calculating the expected plot. Assume for Al $M \sim 27m_p$ where m_p is the proton mass, $\omega_0 \sim 10^{14}$/s, and a lattice constant $a_0 \sim 0.4$ nm.

(e) Discuss if one could potentially use this sort of temperature-dependence to *measure* the temperature of the crystal. Which semiconductors will allow for a decent Debye–Waller "thermometer"?

(10.7) Photons on a ring: Debye–Scherrer X-ray

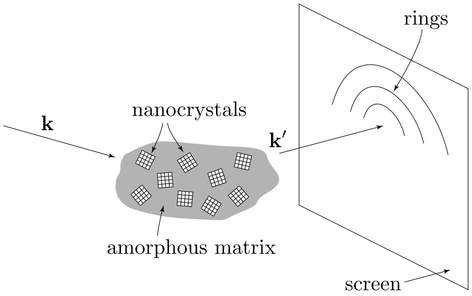


Fig. 10.15 X-ray diffraction from randomly oriented nanocrystals, or polycrystalline films produces ring patterns. This method, first used by Debye and Scherrer in Gottingen, is called powder diffraction.

Semiconductor quantum dot nanocrystals are now popularly used in large-area TV screens (called Q-LEDs TVs). These semiconductor quantum dots control the color of light emitted from them by quantum confinement via their size. They are embedded inside a larger bandgap matrix that is typically amorphous, as indicated in Fig. 10.15. Answer the following questions for such a quantum dot nano-composite material:

(a) Fig. 10.15 shows the scattering of an X-ray beam from nanocrystals embedded in a matrix of an amorphous material. If there was only one nanocrystal, write down the scattering factor, and sketch the pattern of the X-ray photons that would be seen on the screen, and prove that it will be a set of well-defined points.

(b) Now if there are a very large number of nanocrystals whose crystal axes are perfectly aligned with each other, show that the pattern of part (a) will remain the same, but it will only get brighter and sharper.

(b) If there are a very large number of nanocrystals whose crystal axes are aligned along completely random directions, show that the pattern of part (a) will become a ring. So there you have it – this is how you put photons on a ring!

(10.8) This crystal has no atoms! The empty lattice bandstructure

(a) **Impossibility of indirect bandgaps in ideal 1D crystals**: Argue why in an ideal 1D crystal, energy degeneracies (or band crossings) can *only* occur at $k = 0$ or at the Brillouin zone edges, but never at any other k points. Since degeneracy points lead to bandgaps, what does this say about ideal 1D semiconductors: will they be direct or indirect bandgap? What about 2D and 3D crystals?

(b) Now consider the ideal 2D crystal of graphene shown in the next chapter, in Fig. 11.14. Identify the primitive real-space lattice vectors $(\mathbf{a_1}, \mathbf{a_2})$. Find the primitive reciprocal lattice vectors $(\mathbf{b_1}, \mathbf{b_2})$. The general reciprocal lattice vector may be written as $\mathbf{G} = n\mathbf{b_1} + m\mathbf{b_2}$, written more compactly as (n, m).

(c) Starting from the expression for the energy bandstructure of the empty lattice model $\mathcal{E}(\mathbf{k})$ = $\frac{\hbar^2(\mathbf{k}+\mathbf{G})^2}{2m_0}$, draw up a table with expressions of energy bands along the k_x and k_y directions. Associate each band with the respective (n, m) reciprocal lattice vector.

(d) Plot the empty-lattice bandstructure along $(k_x, 0)$ for (n, m): $(0, 0)$ through $(2, 2)$ with all possible $(\pm)$ combinations. Also make a 2D plot of the energy bandstructure. Superimpose and compare with the known bandstructure of graphene and comment.

(e) Do the same as in the previous part, but for 3D silicon (see Fig. 11.17). Use the diamond cubic crystal structure (you have three reciprocal lattice vectors). Plot the energy bands along the typical high-symmetry directions in the k-space. Superimpose and compare with the known bandstructure of silicon and comment.

(10.9) **Photo 51: What is life?**

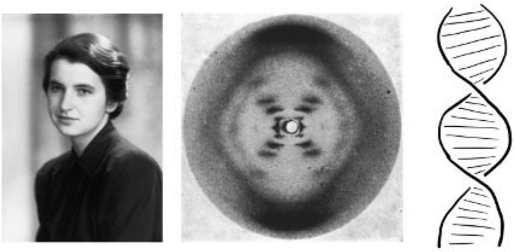


Fig. 10.16 Rosalind Franklin, the famous Photo 51, and the structure of the DNA double-helix.

Fig. 10.16 shows Rosalind Franklin, with a very famous photo she took for the first time: it is so famous that it has a (somewhat unremarkable?) name: Photo 51. Now this is not an ordinary photo, but an X-ray diffraction pattern, and the X-ray beam was going through what at that time was already identified chemically as the blueprint of life itself: the deoxy ribonucleic acid, or DNA. This chemical can self-replicate, and is passed from generation to generation as the mathematical code that determines how our cells will divide, how tall we will be, whether I may be color blind, and so much more.

Now several researchers were trying very hard in the early 1950s to decode the DNA – meaning find its molecular and atomic structure. Among them were Linus Pauling (two Nobel prizes, one in Chemistry and another in Peace), and Francis Crick, James Watson, and Maurice Wilkins (Nobel Prize in Chemistry for the discovery of the structure of the DNA, 1962). Though Rosalind made the key breakthrough, she died in 1958 at age 37 of ovarian cancer before the Nobel Prize for DNA was awarded to Watson, Crick, and Wilkins (though it is not clear if she would have been chosen either!) and is therefore not as well known in popular media as she deserves.

(a) Photo 51 showed for the first time that the X-ray diffraction spectrum exhibits diffraction spots arranged in a X-like shape. What sort of crystal will lead to a diffraction pattern that will have a X-shape? If you can identify the structure and had seen Rosalind's Photo 51 in 1953, you would have decoded the DNA! This is exactly what Watson and Crick were able to do.

(b) Now the sketch on the right in Fig. 10.16 shows a double–helix, which is the correct structure of the DNA molecule. The double helix is formed of two long chain molecules, which are intertwined forming the double helix. There are connections between them, which carry the genetic code, made of the molecules guanine, adenine, thymine, and cytosine, abbreviated as G, A, T, and C (and cleverly used to name movies on genetic engineering such as Gattaca). Identify the lattice and the basis of the DNA molecule. Show that the DNA molecule of this form is a crystal, and should have a diffraction pattern such as seen in Photo 51.

(c) Now we can stop pretending that the DNA is just a molecule, and realize that a long chain of it is periodic, and therefore the electrons in DNA must form bands! If you have identified the lattice and basis in part (b), draw the nearly free–electron bandstructure of the DNA. It turns out that the DNA molecule is a semiconductor, and researchers have been able to measure the current flowing through it – do some research and connect to the bandstructure that you have calculated.

Tight-Binding Bandstructure

11

11.1 Atoms, bonds, and molecules 222
11.2 Bandstructure of 1D, 2D, and 3D crystals 225
11.3 1D, 2D: nanotubes, graphene, BN, MX_2 234
11.4 3D FCC: Si, GaAs 238
11.5 3D wurtzite: GaN, AlN, ZnO 243
11.6 Tight-binding to design new properties 244
11.7 Chapter summary section 245
Further reading 246
Exercises 246

We used the free–electron problem as our "solved problem" $\hat{H}^0|\psi\rangle = E|\psi\rangle$ in the past chapters. Because we know the eigenfunctions $\psi_k(x) = e^{ikx}/\sqrt{L}$ and corresponding eigenvalues $E(k) = \frac{\hbar^2 k^2}{2m_e}$ of the solved problem, we could treat the periodic potential $V(x+a) = V(x)$ of atoms in a crystal as a perturbation to the free–electron potential. In Chapter 8, this method of attack taught us that electrons in a crystal will have bands, gaps, effective masses, and the eigenfunctions will take the form of Bloch functions. The concept of Bloch eigenfunctions and eigenvalues of electron waves in crystals, and its ramifications was developed in its full glory in Chapter 9, and crystal symmetries and the nearly free–electron model of Chapter 10 has prepared us to calculate realistic bandstructures of semiconductors. One of the paths to compute the bandstructure of solids from Bloch theory (Chapter 9) is the tight-binding model. In this chapter, we learn:

- How to extract the properties of molecules and crystalline solids if we know the properties of the constituent atoms, and their geometrical arrangement in real space,
- Develop an intuitive understanding of the tight-binding model for electronic bandstructure of crystals,
- Apply the tight-binding method to calculate and understand the bandstructure and physical properties of various semiconductors, ranging from graphene to silicon to GaN, and
- How to apply the tight-binding method to design and predict new properties such as those that are topological and non-trivial in origin.

The Bloch wavefunction is $\psi_k(x) = e^{ikx}u_k(x)$, where the e^{ikx} part has the imprint of the free–electron wavefunction, and $u_k(x)$ has the imprint of the atomic wavefunction. Chapters 5, 8 and 10 approached the problem from the free–electron viewpoint, excluding the atomic details. The alternative way of looking at the same problem (see Fig. 11.1) is to imagine the atoms to be very far apart to begin with. Because then, since the electrons are bound tightly to them, we know all the possible eigenfunctions (atomic orbitals) and eigenvalues (sharp and discrete atomic electron energies). Fig. 11.2 schematically shows

Fig. 11.1 The American physicist John Slater was one of the earliest to use the tight-binding model for crystalline solids. Introduced the Slater–Koster orbitals, and is known for the Slater determinants for fermion wavefunctions in quantum mechanics. Slater was the PhD advisor of William Shockley, one of the pioneers of semiconductor physics and transistors.

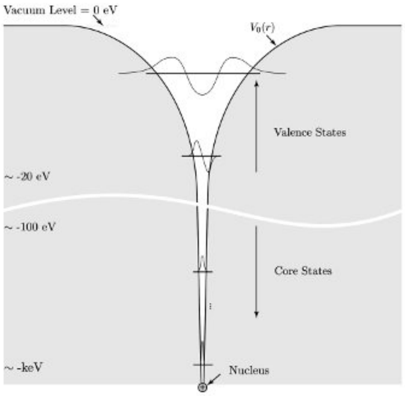


Fig. 11.2 Schematic representation of atomic potential, eigenvalues, and eigenfunctions.

the eigenvalues and eigenfunctions of atomic states, before the formation of chemical bonding. For atoms that form semiconductor crystals (such as silicon), the deepest electron energy states due to the potential $V_0(r)$ (such as the $|1s\rangle$ orbital) are $\sim$ 1000 eV below the vacuum level. Because electrons in the deep energy levels have strongly localized wavefunctions due to enormous potential barriers that confine them, they are unable to participate in chemical bonding. We will call them the **core states**. The core state electrons can be ejected by the photoelectric effect from the semiconductor by photons of high energy ($\sim$ keV) such as X-rays. The X-ray Photoelectron Spectroscopy (XPS) technique provides a fingerprint of the chemical bonding nature of the semiconductor. On the other hand, electrons in the top-most shallow energy states are called **valence states** because they have much smaller barriers and are shared between atoms to form chemical bonds, as indicated in Fig. 11.3. Assuming the core electrons to be a "part of the nucleus" so to speak, to understand the crystal bandstructure, we focus on the valence electrons, which lie typically within $\sim$ 20 eV below the vacuum level for most semiconductors and consist primarily of $|s\rangle$ and $|p\rangle$ orbitals of the outermost atomic shells.

11.1 Atoms, bonds, and molecules

Consider electrons in the outermost $|s\rangle$-orbital states with eigenfunctions $\langle r|s\rangle = \psi_s(r)$, and sharp eigenvalues E_s which are obtained from the atomic Hamiltonian $\hat{H}^0|s\rangle = E_s|s\rangle$ with potential $V_0(r)$. This is shown in Fig. 11.3. We can consider this set of eigenfunctions and eigenvalues to be our exactly "solved problem". Now to build a crystal, we start by asking: if I bring two such atoms close to each other, how do the eigenfunctions and eigenvalues change? Note how this is a different perspective from the free-electron case. Instead of electrons being free, the exactly solved problem is one where the electron is "tightly bound" to the atom, and only by interaction with nearby orbitals can break free and move. The eigenfunction $\psi_s(r)$ starts out localized, and will evolve into extended Bloch states $\psi_k(r) = \sum_R e^{ikR}\psi_s(r-R)$ when we form a crystal. The eigenvalues E_s start out sharp, and will evolve into bands $E(k) = E_s - 2V_{ss\sigma}\cos(kR)$ once the crystal is formed. The new eigenfunctions and eigenvalues represent the situation when two atoms bond to form a molecule. As is seen in Fig. 11.3, the total potential $V(r)$ seen by an electron between the two nuclei is *lowered* compared to the potential $V_0(r)$ due to a single nucleus. The Hamiltonian is $\hat{H} = \hat{H}^0 + W(r)$, where $W(r) = V(r) - V_0(r)$ is the potential lowering, seen as a perturbation to the atomic state. As a result, the electron wavefunction can penetrate a bit more into the barrier, overlapping with the similar part leaked from the other atom, and is now shared between the nuclei. This is the genesis of the **chemical bond**.

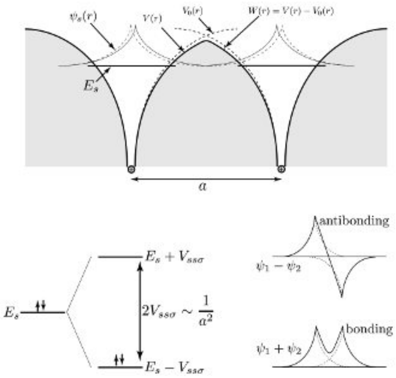


Fig. 11.3 Chemical bonding energies and eigenfunctions. Though the eigenvalues $E_s \pm V_{ss\sigma}$ are shown equally split around E_s, a more accurate calculation that does not neglect the wavefunction overlaps shifts both energies slightly upwards. The antibonding energy is further above E_s than the bonding energy is below, so their sum is slightly *higher* than $2E_s$. This is the reason why H_2 molecules exist, yet there is no He_2 molecule.

Though the above perspective is different from the free–electron case, the method of solving this perturbation problem is not. To find the new eigenvalues and eigenfunctions of the molecule compared to the atom, we apply the matrix method developed in Chapter 7. The **input** to the problem is the distance between the atoms. Because the $|s\rangle$ orbital wavefunctions are centered around the nucleus of each atom, the new eigenstates, by the expansion principle, are $|\psi\rangle = a_1|s_1\rangle + a_2|s_2\rangle$. Here $\langle r|s_1\rangle = \psi_s(r - r_1) = \phi_1$ and $\langle r|s_2\rangle = \psi_s(r - r_2) = \phi_2$: they are exactly the same eigenfunction, but with their centers separated by $a = r_1 - r_2$, the distance between the nuclei. The Schrödinger equation is $\hat{H}|\psi\rangle = E|\psi\rangle$, which written out explicitly as projection on states $\langle s_1|$ and $\langle s_2|$ is[1]

$$[\hat{H}^0 + W] = \begin{array}{c c} & \begin{array}{cc} |s_1\rangle & |s_2\rangle \end{array} \\ \begin{array}{c} \langle s_1| \\ \langle s_2| \end{array} & \begin{pmatrix} E_s & -V_{ss\sigma} \\ -V_{ss\sigma} & E_s \end{pmatrix} \end{array}, \tag{11.1}$$

where we have written $H_{11} = \langle s_1|\hat{H}^0 + W|s_1\rangle = E_s + \langle s_1|W|s_1\rangle \approx E_s$ if the diagonal perturbation is very small compared to the on-site potential. It can be included if the diagonal shift is comparable to E_s. The cross term $\langle s_2|\hat{H}^0 + W|s_1\rangle = \langle s_2|\hat{H}^0|s_1\rangle + \langle s_2|W|s_1\rangle = E_s\langle s_1|s_2\rangle + \langle s_2|W|s_1\rangle \approx \langle s_2|W|s_1\rangle = -V_{ss\sigma}$, the lowering of energy because of hopping between s-orbitals. Since $W(r)$ is negative, we choose $V_{ss\sigma}$ to be the magnitude of the matrix element. Because $E_s\langle s_1|s_2\rangle << \langle s_2|W|s_1\rangle$, it is assumed to be zero. The reason for the small value of $\langle s_1|s_2\rangle$ and the negative sign of the cross-matrix term can be seen in Fig. 11.3.

The molecular tight-binding problem is solved in its entirety by finding the eigenvalues and eigenfunctions of Equation 11.1. The eigenvalues and eigenfunctions are

$$\begin{aligned} E_+ &= E_s + V_{ss\sigma}, \text{ and } \psi_+(r) = \frac{1}{\sqrt{2}}[\phi_1 - \phi_2] \\ E_- &= E_s - V_{ss\sigma}, \text{ and } \psi_-(r) = \frac{1}{\sqrt{2}}[\phi_1 + \phi_2], \end{aligned} \tag{11.2}$$

with corresponding eigenvectors $a_- = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}$ and $a_+ = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}$. From Fig. 11.3, the electron wavefunction peaks between the nuclei for the lower energy bonding orbital $\psi_-(r)$ – this is the **bonding state**, and the increase in $|\psi(r)|^2$ between the nuclei compared to the free atomic state implies an increased probability of electrons being found in-between, or shared between the atoms – this is the **chemical bond**. For $\psi_+(r)$, the wavefunction goes to zero at the center, indicating an **anti-bonding state**. This is similar to the ground state and first excited state of the particle in a box. If each atom had one electron in the outermost orbital before bonding, the net energy of the two independent atoms was $2E_s$. After chemical bonding, as Fig. 11.3 shows, the energy is $2E_s - 2V_{ss\sigma}$. The system has lowered[2] its energy by $2V_{ss\sigma}$.

Exactly which orbitals participate in chemical bonding depends, of course, on the atoms. In chemistry the most important orbitals of the

[1] The physical meaning of the matrix elements is the following. If the electrons stay put on an atomic site, their "on-site" energy is E_s. If they hop over to the next atom, they can *lower* their energy by $-V_{ss\sigma}$. So the off-diagonal terms are referred to as the "hopping energy". If the electron can spread out in space, it can assume longer wavelengths, and therefore it can lower its energy. Just as the particle in a box, the hopping energy $V_{ss\sigma} \sim \frac{1}{a^2}$ decreases as the square of the distance between the atoms, which roughly is the length over which the electron wavefunction can delocalize. Such Hamiltonians are sometimes written as $\hat{H} = E_s c_1^\dagger c_1 + E_s c_2^\dagger c_2 - V_{ss\sigma} c_2^\dagger c_1 - V_{ss\sigma} c_1^\dagger c_2$, where c_1 is annihilation operator for an electron in site 1, and $c_1^\dagger$ is the creation operation for the same site, and corresponding operators for site 2.

[2] This is the reason why hydrogen, oxygen, nitrogen, and their combinations form molecules rather than stay in their atomic states. For example, the energy required to break the molecule N_2 into two N atoms is precisely $2V_{ss\sigma}$, giving us an experimental method to directly measure the hopping energies.

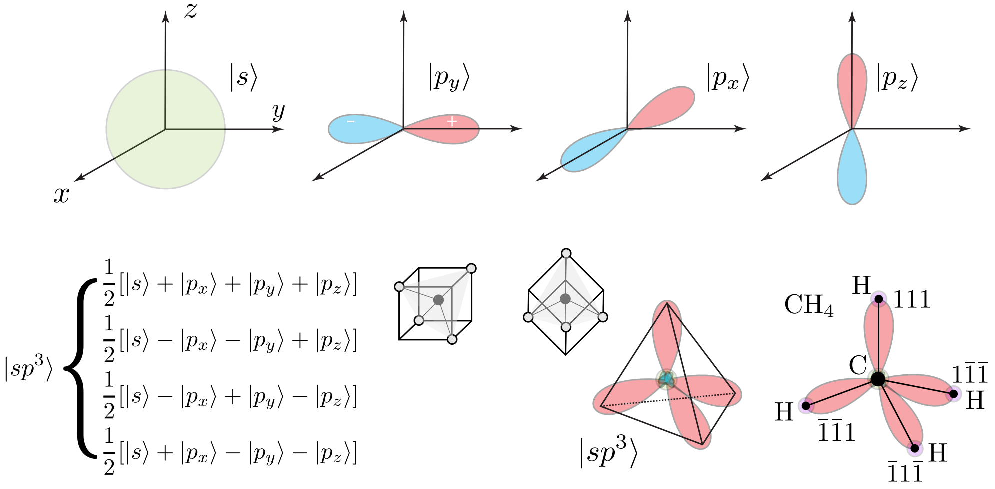

z
$|s\rangle$
y
x
$|p_y\rangle$
−
+
$|p_x\rangle$
$|p_z\rangle$
$|sp^3\rangle$
$\frac{1}{2}[|s\rangle + |p_x\rangle + |p_y\rangle + |p_z\rangle]$
$\frac{1}{2}[|s\rangle - |p_x\rangle - |p_y\rangle + |p_z\rangle]$
$\frac{1}{2}[|s\rangle - |p_x\rangle + |p_y\rangle - |p_z\rangle]$
$\frac{1}{2}[|s\rangle + |p_x\rangle - |p_y\rangle - |p_z\rangle]$
$|sp^3\rangle$
CH_4
H
111
C
$1\bar{1}\bar{1}$
H
H
$\bar{1}\bar{1}1$
H
$\bar{1}1\bar{1}$

Fig. 11.4 s and p atomic orbitals, and the linear combination of atomic orbitals (LCAO) to form sp^3 bonds. The linear combinations resulting in the sp^3 bond are shown. The sp^3 bond is tetrahedral, and so is the CH_4 molecule that results.

Fig. 11.5 Linus Pauling made early and decisive contributions to the quantum mechanical understanding of the chemical bond, and the underlying mechanisms of Chemistry. He identified the sp^3 hybridization of the methane molecule for the first time in 1931. He was awarded two Nobel Prizes, one in Chemistry and the other in Peace.

two atoms participating in chemical bonding are called ”frontier orbitals”. In typical semiconductors, the outermost $|s\rangle$ and $|p\rangle$ orbitals participate, though $|d\rangle$ orbitals are of increasing interest to host additional physical properties. Fig. 11.4 shows the shapes of $|s\rangle$ and $|p\rangle$ orbitals. The $|s\rangle$ orbital wavefunction is spherically symmetric around the nucleus it is tied to, whereas the $|p\rangle$ orbital has dumbbell shaped lobes; the wavefunction is odd around the origin: its sign flips at the nucleus. The three basis sets for the $|p\rangle$ orbitals align along mutually orthogonal directions, along the x-, y-, and z-axes. Consider the methane molecule CH_4. The $|1s\rangle$ orbitals of the four H atoms are occupied as $1s^1$ each of energy eigenvelaue E_{1s}. The six electrons of the C atoms fill the orbitals in the configuration $1s^2 2s^2 2p^2$. Of these, the $1s^2$ electrons form the deep-level core states and do not participate in bonding. There are four bonds formed between eight electrons: between the $2s^2 2p^2$ electrons of C, and the four $1s^1$ electrons of each H atom. The methane molecule is known from experiments to be a tetrahedron (as seen in Fig. 11.4) – which means that the tetrahedral chemical bonds must be linear combinations of the participating chemical bonds. In 1931, soon after the development of quantum mechanics, Pauling (Fig 11.5) found that the tetrahedral bonds of methane require a linear combination of 1 s and 3 p orbitals, called the sp^3 chemical bond. A way to visualize the chemical bonding is:

$$C : 2s^2 2p_x^1 2p_y^1 \rightarrow C : \underbrace{2s^1 2p_x^1 2p_y^1 2p_z^1}_{sp^3} \rightarrow \underbrace{sp^3}_{C} + 4 \times [\ \underbrace{1s^2}_{\text{H atom}}\ ] = CH_4. \tag{11.3}$$

The four linear combinations to form sp^3 bonds are written out in Fig. 11.4, next to a sketch of the four lobes of the tetrahedral bonds. Note that it costs energy for the four electrons of carbon atom to hybridize to form sp^3, rather than stay in its lowest energy atomic state $2s^2 2p_x^1 2p_y^1$. After all promoting an electron from the 2s orbital to a 2p orbital must cost energy, since the 2p orbital has a higher energy. So why does C hybridize to form CH_4? Because by hybridizing just so, it is able to form chemical bonds with four H atoms at the tips of the tetrahedra, and the energy lowering due to this bonding *more than recovers* the cost of electron promotion[3]. So the energy eigenvalues of the molecule (Fig. 11.3) tell us the most important thing about bandstructures: that when wavefunctions add constructively, the electron is delocalized over the atoms, and the corresponding bonding energy eigenvalue is lower than the unbonded energy. If they add destructively, the corresponding antibonding energy eigenvalue is higher than the unbonded energy. Now we are ready to understand the tight-binding electron bandstructure, or $E(k)$ diagrams of 1D, 2D, and 3D crystals.

11.2 Bandstructure of 1D, 2D, and 3D crystals

Consider an atom with a valence s-orbital of energy E_s. Two of them go through bonding and antibonding, with energies[4] $E_s \pm V_{ss\sigma}$ (Fig. 11.6 (a)). Consider now a chain of N of them lined up (Fig. 11.6 (b), s-bands) in a periodic lattice of lattice constant a. Labeling the on-site atomic orbitals as $\phi_n = \psi_s(r - na)$, the bonding state is the sum of all orbitals $\psi = \phi_0 + \phi_1 + \phi_2 + ...$ This wavefunction *must*[5] satisfy the Bloch theorem $\psi_k = \sum_n e^{ikna}\phi_n$, where e^{ikna} are the coefficients of a linear combination. It is evident that the bonding state is achieved when $k = 0$, $e^{ikna} = e^0 = +1$ for all n, and all wavefunctions add in phase, $\psi_{k=0} = \sum_n e^{i0}\phi_n = \phi_0 + \phi_1 + \phi_2 + \phi_3 + ...$ The energy of this state must be quite a bit lower compared to E_s because the electron in this k-state is *delocalized* over a large number N of atoms. The $k = 0$ point is the Γ-point in the $E(k)$ diagram, as shown in Fig. 11.6 (b).

For $k = \pi/a$, the Bloch sum becomes $\psi_{k=\pi/a} = \sum_n e^{in\pi}\phi_n = \phi_0 - \phi_1 + \phi_2 - \phi_3 + ...$ We identify it as the crystal-analogue of the antibonding linear combination of the molecule, and also realize that this Bloch state must have the highest energy. If there are N atoms[6], $L = Na$, and $e^{ikL} = 1$, implying that k can only take values $k_n = \frac{2\pi}{L}n$ where $n = 0, 1, ..., N-1$, and the phases for $k = +\frac{\pi}{a}$ and $k = -\frac{\pi}{a}$ are *identical*, meaning they are exactly the same Bloch state[7]. This is vividly seen in Fig. 11.6. This means that the k-axis is a ring!

There are exactly N k-points in the interval $-\frac{\pi}{a} < k \leq +\frac{\pi}{a}$, meaning there are exactly N energy eigenvalues, distributed in some functional form between $E(k = 0)$ and $E(k = \frac{\pi}{a})$. This is no surprise – we started with N eigenvalues all at energy E_s, and we ended with N eigenvalues, just that they are now spread out in energy $E(k)$ around E_s. We identify $k = \pm\frac{\pi}{a}$ as the familiar Brillouin-zone edges, but with added insight about the nature of the Bloch wavefunctions there. The functional form $E(k)$ is the band that the tight-binding method will

---

[3] The answers to many vexing *why* questions for chemical bonding may be obtained by realizing that electrons and atoms take a path to the lowest *final* energy, even if they have to go over an *intermediate* barrier. The molecule can lower its energy compared to the original unbonded atoms if the number of electrons are such that when the new energy eigenvalues are filled following the Pauli exclusion principle, the total energy of the molecule is lower than the energy of the unbonded atoms.

[4] Neglecting the uneven splitting around the original energy.

[5] Why? Because the electrons experience a periodic potential $V(r + a) = V(r)$.

[6] Assume they are on a circular ring – you must be very familiar with this procedure by now!

[7] They are different by a sign, $\psi_{k=\pi/a} = -\psi_{k=-\pi/a}$, but $|\psi_{k=\pi/a}|^2 = |\psi_{k=-\pi/a}|^2$.

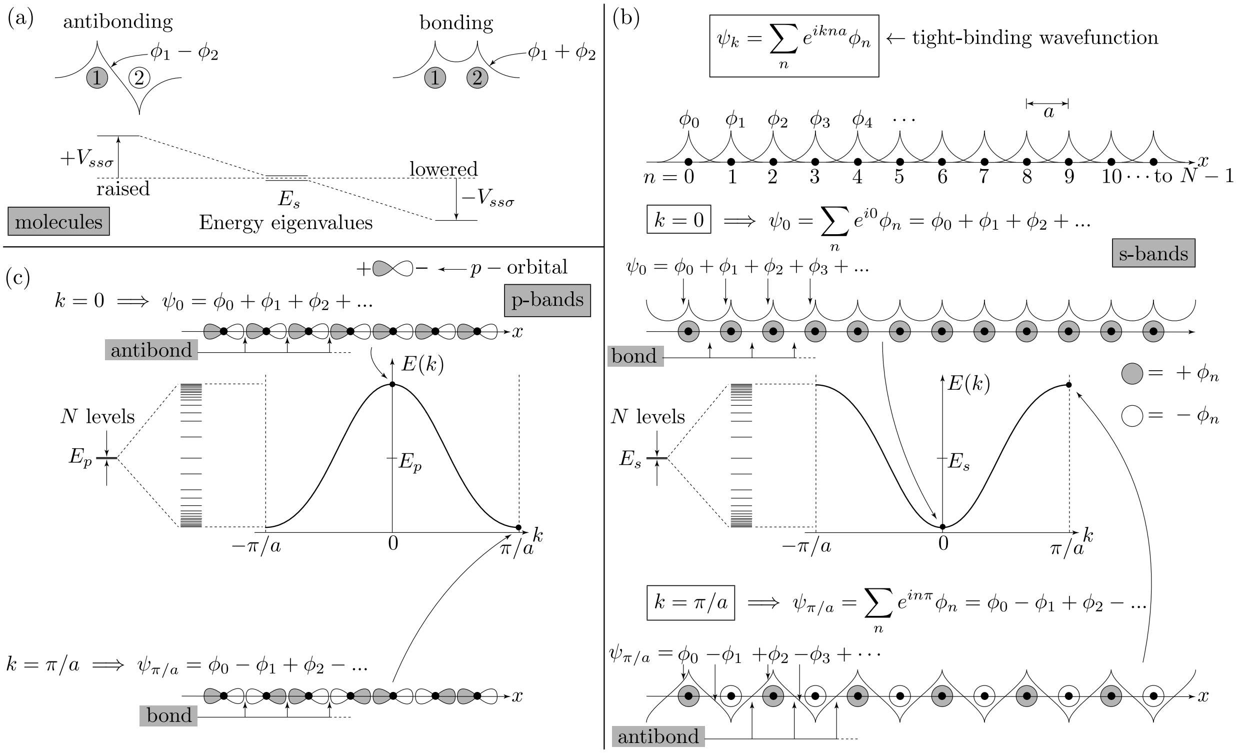

(a)
antibonding
$\phi_1 - \phi_2$
1
2
bonding
$\phi_1 + \phi_2$
$+V_{ss\sigma}$
raised
E_s
lowered
$-V_{ss\sigma}$
molecules
Energy eigenvalues
(b)
$\psi_k = \sum_n e^{ikna}\phi_n$ ← tight-binding wavefunction
ϕ_0 ϕ_1 ϕ_2 ϕ_3 ϕ_4 ···
a
x
$n = 0$ 1 2 3 4 5 6 7 8 9 10 ··· to $N-1$
$k = 0$ $\Longrightarrow$ $\psi_0 = \sum_n e^{i0}\phi_n = \phi_0 + \phi_1 + \phi_2 + ...$
s-bands
$\psi_0 = \phi_0 + \phi_1 + \phi_2 + \phi_3 + ...$
bond
$E(k)$
$= +\phi_n$
$= -\phi_n$
N levels
E_s
$-\pi/a$
0
π/a
k
$k = \pi/a$ $\Longrightarrow$ $\psi_{\pi/a} = \sum_n e^{in\pi}\phi_n = \phi_0 - \phi_1 + \phi_2 - ...$
$\psi_{\pi/a} = \phi_0 - \phi_1 + \phi_2 - \phi_3 + \cdots$
antibond
(c)
$+$ $-$ ← p − orbital
$k = 0 \Longrightarrow \psi_0 = \phi_0 + \phi_1 + \phi_2 + ...$
p-bands
antibond
N levels
E_p
$E(k)$
E_p
$-\pi/a$
0
π/a
k
$k = \pi/a \Longrightarrow \psi_{\pi/a} = \phi_0 - \phi_1 + \phi_2 - ...$
bond

Fig. 11.6 Formation of molecules, s-bands, and p-bands.

now give us. We can predict a few things about this band: if the hopping energy $V_{ss\sigma}$ is small, the energy bandwidth of this band will be small, with the band collapsing to its atomic sharp eigenvalue of E_s as $V_{ss\sigma} \to 0$. Since the density of states is the number of states within an energy interval, a small $V_{ss\sigma}$ will increase the density of states, and also result in large effective masses. On the other hand, if $V_{ss\sigma}$ is large, the band width will be large, and the curvature at the top and bottom will increase: this will lead to small effective masses $m^\star \sim \frac{1}{V_{ss\sigma}}$ at the band edges. These qualitative characteristics will be soon borne out in most semiconductors.

Now consider instead of s orbitals, a crystal in which the bonding is between p orbitals of unbonded energy E_p, as shown in Fig. 11.6 (c) (p-bands). The Bloch function[8] corresponding to $k = 0$ is now the *antibonding* state, unlike the s-band, and the $k = \pm\frac{\pi}{a}$ states are the bonding states. The $k = 0$ and $k = \pm\frac{\pi}{a}$ states have swapped roles! It is clear why it must be so: when the p-orbital wavefunctions are added in phase $\psi_{k=0} = \phi_0 + \phi_1 + \phi_2 + ...$, because the p-orbital is antisymmetric around each nuclei, there is a node between the nuclei and an antibonding state is formed, with higher energy than E_p. But if we flip alternate p-orbitals around the nuclei and take this linear combination

---

[8] The tight-binding Bloch functions are sometimes also called LCAO wavefunctions (see Figure 11.4).

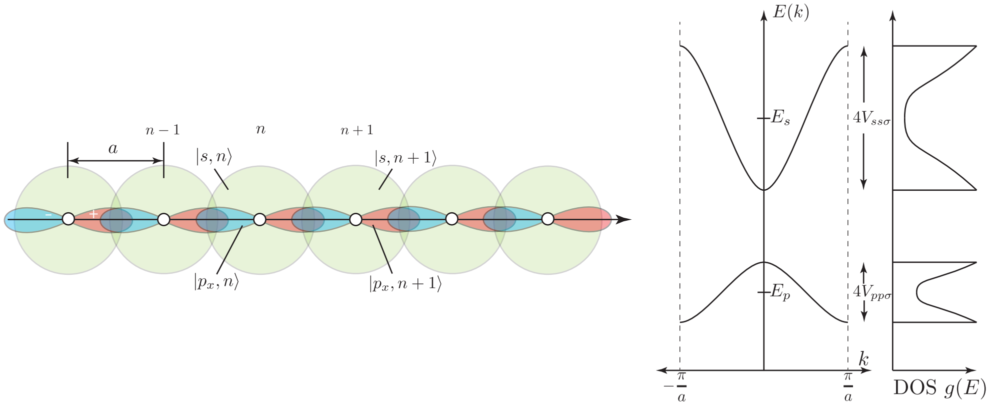


Fig. 11.7 Overlaps between nearest neighbor s- and p-orbitals of a 1D crystal. The s-orbital is spherical, and hence symmetric along all axes. The p-orbitals are antisymmetric or odd along the direction they are oriented – i.e., the p_x orbital has two lobes – one positive, and the other negative.

$\psi_{k=+\pi/a} = \phi_0 - \phi_1 + \phi_2 - \phi_3 ...$, the wavefunctions add between the nuclei, forming a bonding state. This leads to a p-band that looks like the s-band flipped upside down: the $k = 0$ or Γ-point of the Brillouin zone is where the *minimum of the s-orbital band* of eigenvalues, and the *maximum of the p-orbital band* eigenvalues is reached. For each of the bands, there are N states in the BZ. Each band can hold a maximum of $2N$ electrons. If each atom contributes one electron, the bands are half full, and if they contribute two electrons each, they are completely filled.

Now consider Fig. 11.7, for which one s-orbital and one p_x-orbital of each atom participate in chemical bonding. In this case, the overlaps, or hops possible are $s-s$, $p-p$, and also $s-p$ between nearest neighbors. The corresponding hopping energies are $V_{ss\sigma}$, $V_{pp\sigma}$, and $V_{sp\sigma}$. Because we started with two orbitals per atom, each orbital leads to a band, and the bandstructure will thus contain two bands. If we assume that $V_{sp\sigma} = 0$, then the bands will be a s-band, and a p-band of Fig. 11.6 together, as seen in Fig. 11.7. Depending on the values of E_s and E_p, and the hopping energies $V_{ss\sigma}$ and $V_{pp\sigma}$, the bands may or may not overlap. It may also be the case that one band is completely contained within the other[9]! Note that for each k, there are now two bands: one from s-orbitals, and the other from p-orbitals. If we now allow hopping from s to nearest neighbor p orbitals, then there are still two energy eigenvalues for each k, but we cannot cleanly say if it is s or p: this is the meaning of a non-zero $V_{sp\sigma}$.

Before discussing the tight-binding procedure, it is useful to develop an intuitive feeling for what new things can happen in 2D and 3D that are not seen in 1D. Fig. 11.8 shows the evolution of bands in 2D. For simplicity we have assumed a square lattice of side a. The $\mathbf{k} = (k_x, k_y)$-space and the $\mathbf{r} = (x, y)$-spaces are 2D now, the atomic orbitals are

[9] This phenomena is responsible for the sudden increase in electrical and thermal conductivity going from Ni to Cu in the periodic table: a narrow d-band is completely contained within a broad s-band.

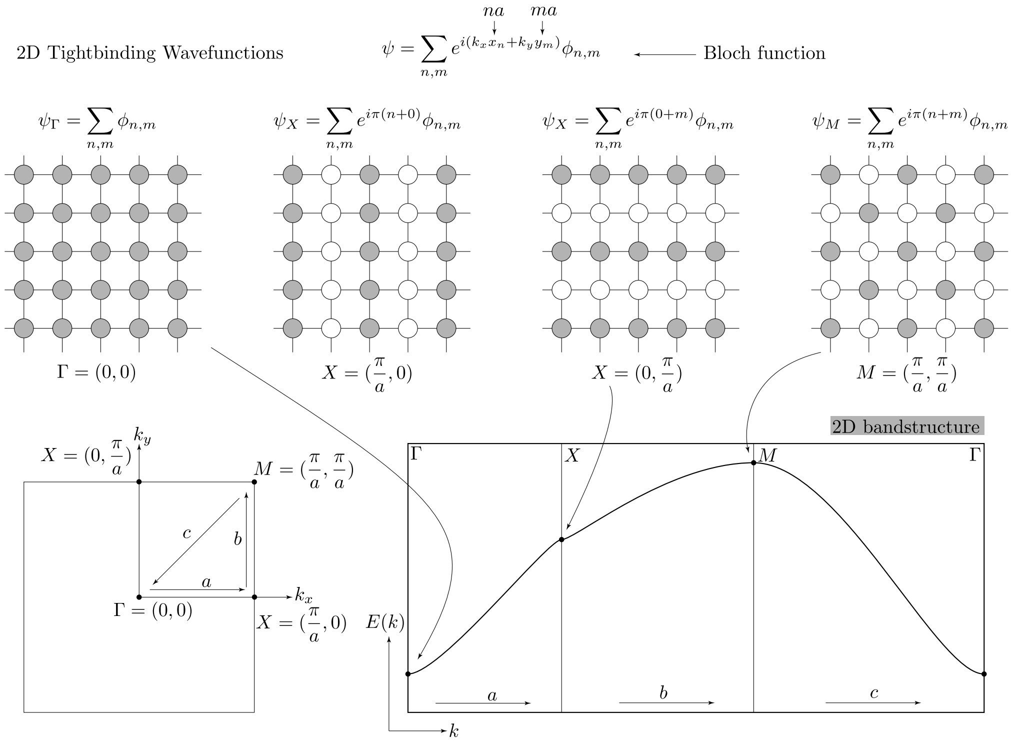

2D Tightbinding Wavefunctions
na
ma
$\psi = \sum_{n,m} e^{i(k_x x_n + k_y y_m)}\phi_{n,m}$
Bloch function
$\psi_\Gamma = \sum_{n,m} \phi_{n,m}$
$\psi_X = \sum_{n,m} e^{i\pi(n+0)}\phi_{n,m}$
$\psi_X = \sum_{n,m} e^{i\pi(0+m)}\phi_{n,m}$
$\psi_M = \sum_{n,m} e^{i\pi(n+m)}\phi_{n,m}$
$\Gamma = (0,0)$
$X = (\frac{\pi}{a}, 0)$
$X = (0, \frac{\pi}{a})$
$M = (\frac{\pi}{a}, \frac{\pi}{a})$
2D bandstructure
$X = (0, \frac{\pi}{a})$
k_y
$M = (\frac{\pi}{a}, \frac{\pi}{a})$
c
b
a
k_x
$\Gamma = (0,0)$
$X = (\frac{\pi}{a}, 0)$
$E(k)$
Γ
X
M
Γ
a
b
c
k

Fig. 11.8 Tightbinding bandstructure in two dimensions.

$\phi_{n,m}(r)$ centered at each site $\mathbf{r} = (na, ma)$, and the Bloch function is

$$\psi_{\mathbf{k}} = \sum_{n,m} e^{i\mathbf{k}\cdot\mathbf{r}}\phi_{n,m} = \sum_{n,m} e^{i(nk_x+mk_y)a}\phi_{n,m} \underbrace{=}_{(k_x=k_y=0)} \psi_\Gamma = \sum \phi_{n,m}. \quad (11.4)$$

We see the Bloch wavefunction for the state at the $\Gamma = (k_x, k_y) = (0,0)$ point is a simple sum of all orbitals in phase. If the bonding is between s-orbitals, this will be the lowest energy in the band because it will be the bonding state. Fig. 11.8 represents a +ve ϕ_n as a filled circle, and a negative ϕ_n as empty; the Γ state is shown as the sum $\psi_\Gamma = \phi_{0,0} + \phi_{0,1} + ...$ In 1D, there were only two special points in the k-space: $k = 0$ and $k = \frac{\pi}{a}$. In a 2D square lattice, points of high symmetry are labeled in Fig. 11.8 as Γ, $X = (\frac{\pi}{a}, 0)$, and $M = (\frac{\pi}{a}, \frac{\pi}{a})$. X and M are points at the BZ edges, obtained by the perpendicular bisectors of the lines connecting the Γ point to the reciprocal lattice vectors $n(\frac{2\pi}{a}, 0) + m(0, \frac{2\pi}{a})$. The Bloch function for the M point $\psi_M = \sum_{n,m} e^{i(n+m)\pi}\phi_{n,m}$ is then a checkerboard pattern that has destructive interference, or nodes between every two nearest neighbors: this the most antibonding state of all, and must have the highest energy. Similarly, the X points must have phases that go either along the x- or y-axes, with nodal lines between atomic lines. The X-state eigenvalue then must be intermediate between the Γ and M points.

The Bloch phases $e^{i\mathbf{k}\cdot\mathbf{r}}$ are seen vividly as the waves of the colors: for all intermediate points in the k-space, the colors will be intermediate and the phases have a vector direction due to $\mathbf{k}\cdot\mathbf{r}$. The energy bandstructure could be plotted as a plane surface vs (k_x, k_y), but typically it is difficult to see the bandstructure details in such a 3D plot. It is more common to plot the energy eigenvalues along a closed loop connecting the high symmetry points in the k-space, as shown for the line $\Gamma \rightarrow X \rightarrow M \rightarrow \Gamma$ in Fig. 11.8.

In three dimensions $\mathbf{k} = (k_x, k_y, k_z)$ and $\mathbf{r} = (x, y, z)$, and one can plot similar figures of orbitals and their respective signs in the linear combinations forming the Bloch states. We had identified the k-space paths of high symmetry and the BZs for 2D, and 3D FCC and HCP crystals in Chapter 10 (especially see Figures 10.10 and 10.12): it is time to use them!

We now outline a simplified general tightbinding bandstructure calculation procedure for all dimensions[10]. We write down a general tight-binding model from Bloch linear combinations of orbitals $|n\rangle = \phi_n = \phi(r - R_n)$ that are localized at each lattice point to coax out the eigenvalues $E(k)$ and corresponding eigenfunctions ψ_k. The Bloch state is

$$|\psi_{\mathbf{k}}\rangle = \sum_{n=1}^{N} \frac{e^{i\mathbf{k}\cdot\mathbf{R}_n}}{\sqrt{N}} |n\rangle, \tag{11.5}$$

where $\psi_{\mathbf{k}}(\mathbf{r}) = \langle \mathbf{r}|\psi_{\mathbf{k}}\rangle$, $\mathbf{R}_n$ are the real space lattice vectors, N is the total number of lattice sites, and we start with just one orbital per lattice site. We feed this vector into the Schrödinger equation, cancel $\sqrt{N}$ from each side, and get the relation for energy eigenvalues $E(\mathbf{k})$ for each value of $\mathbf{k}$:

$$\hat{H} \sum_{n=1}^{N} e^{i\mathbf{k}\cdot\mathbf{R}_n} |n\rangle = E(\mathbf{k}) \sum_{n=1}^{N} e^{i\mathbf{k}\cdot\mathbf{R}_n} |n\rangle. \tag{11.6}$$

We note[11] that $\hat{H}$ does not affect $e^{i\mathbf{k}\cdot\mathbf{R}_n}$, it acts only on $|n\rangle$. Next we project both vectors on to the vector $\sum_{m=1}^{N} e^{i\mathbf{k}\cdot\mathbf{R}_m}\langle m|$, and rearrange to get the central result of the tight-binding energy bandstructure $E(\mathbf{k})$:

$$\boxed{E(\mathbf{k}) = \frac{\sum_{n,m=1}^{N} e^{i\mathbf{k}\cdot(\mathbf{R}_n - \mathbf{R}_m)} \langle m|\hat{H}|n\rangle}{\sum_{n,m=1}^{N} e^{i\mathbf{k}\cdot(\mathbf{R}_n - \mathbf{R}_m)} \langle m|n\rangle}.} \tag{11.7}$$

The meaning of this tight-binding bandstructure is delightfully simple. The numerator is a sum over the matrix elements $\langle n|\hat{H}|m\rangle$ that are a measure of the energy for an electron to hop from an atom at lattice site n to site m, including $n = m$, which is the on-site energy eigenvalue. The sum has the Bloch coefficients, or phase factors $e^{i\mathbf{k}\cdot(\mathbf{R}_n - \mathbf{R}_m)}$, whose argument is simply the $\mathbf{k}$ wavevector dotted with the vector $(\mathbf{R}_n - \mathbf{R}_m)$ connecting the two lattice points. We are now familiar with the nature of these phase factors, and how they lead to bonding or antibonding states depending on $\mathbf{k}$. The denominator has the direct overlap of the orbital wavefunctions centered at n and m, modulated by the same Bloch phase factor as the numerator.

---

[10] We keep the math to a minimum using the simplest model, and expand later.

[11] This is because $\hat{H}$ acts on the space coordinate $\mathbf{r}$, but $\mathbf{R}_n$ is a constant, and not the variable $\mathbf{r}$. In real-space coordinates, $\mathbf{k}$ is simply a constant: a set of eigenfunctions and eigenvalues exist for each $\mathbf{k}$. $\hat{H}$ acts only on the atomic orbital $\mathbf{r}$ of $\phi(r - R_n)$, which results from $\langle r|n\rangle$.

There are N^2 terms in the double sum. Out of these, N "diagonal" terms have $n = m$, for which we have $e^{i\mathbf{k}\cdot(\mathbf{R}_m-\mathbf{R}_n)} = e^{i\mathbf{k}\cdot(0)} = 1$ and the diagonal matrix elements are all equal: $\langle n|\hat{H}|n\rangle = E_0$. This energy is slightly *lower* than the original "atomic" orbital energy because each electron orbital also sees nearby atomic potentials. In the denominator, the diagonal sum gives just $\sum e^{i\mathbf{k}\cdot(0)}\langle n|n\rangle = N$. Let us now look at the rest $N^2 - N$ off-diagonal terms in the numerator and denominator. Considering a 1D lattice of lattice constant a, for the first nearest neighbors, N terms in the numerator take the form $\sum_{n=1}^{N}(e^{+ika}\langle n|\hat{H}|n+1\rangle + e^{-ika}\langle n-1|\hat{H}|n\rangle) = -2Nt_1\cos(ka)$, where $t_1 = \langle n|\hat{H}|n+1\rangle$ is the hopping integral. Similarly, the denominator gives the sum of the N first nearest neighbor terms as $+2Ns_1\cos(ka)$, where $s_1 = \langle n|n+1\rangle$ is clearly very small because of decaying wavefunctions that are tightly bound to the lattice sites. There are N more terms for the second nearest neighbors characterized by the hopping integral $t_2 = \langle n|\hat{H}|n+2\rangle$ and the overlap integral $\langle n|n+2\rangle = s_2$. And then for the third nearest neighbor, and so on... It is intuitively clear that the successive terms t_n and s_n decay fast. The number of lattice sites N cancels from the numerator and denominator. Now we can write the expression for the 1D tight-binding bandstructure as:

$$E(k) = \frac{E_0 - 2t_1\cos(ka) - 2t_2\cos(2ka) - 2t_3\cos(3ka)...}{1 + 2s_1\cos(ka) + 2s_2\cos(2ka) + 2s_3\cos(3ka)...} \approx E_0 - 2t_1\cos(ka). \tag{11.8}$$

The approximation $E(k) \approx E_0 - 2t_1\cos(ka)$ is valid if the nearest neighbor hopping terms are much larger in magnitude than the others, $t_1 >> t_2, t_3, ...$, and the wavefunction overlaps are much smaller than unity, $s_1, s_2, ... << 1$.

What if there are two, three, or more orbitals at each lattice site? Calling the orbitals at each site $a, b, c, ...$, the starting Bloch function for the tightbinding state becomes

$$|\psi_{\mathbf{k}}\rangle = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}[c_a e^{i\mathbf{k}\cdot\mathbf{R}_{na}}|na\rangle + c_b e^{i\mathbf{k}\cdot\mathbf{R}_{nb}}|nb\rangle + ...], \tag{11.9}$$

which leads to **as many bands as there are orbitals attached to each lattice site**. The on-site wavefunction is also a linear combination over various orbitals with coefficients $c_a, c_b, ...$ Finally, to each lattice site is attached the basis set of atoms, which can be one, or more often two, three, or more. The Bloch function should then include a summation over the basis set of atoms as well:

$$|\psi_{\mathbf{k}}\rangle = \frac{1}{\sqrt{N}}\Big([c_{1a}e^{i\mathbf{k}\cdot\mathbf{R}_{1a}}|1a\rangle + c_{1b}e^{i\mathbf{k}\cdot\mathbf{R}_{1b}}|1b\rangle + ...]_{\text{atom 1}} + [c_{2a}e^{i\mathbf{k}\cdot\mathbf{R}_{2a}}|2a\rangle + c_{2b}e^{i\mathbf{k}\cdot\mathbf{R}_{2b}}|2b\rangle + ...]_{\text{atom 2}} + ...\Big), \tag{11.10}$$

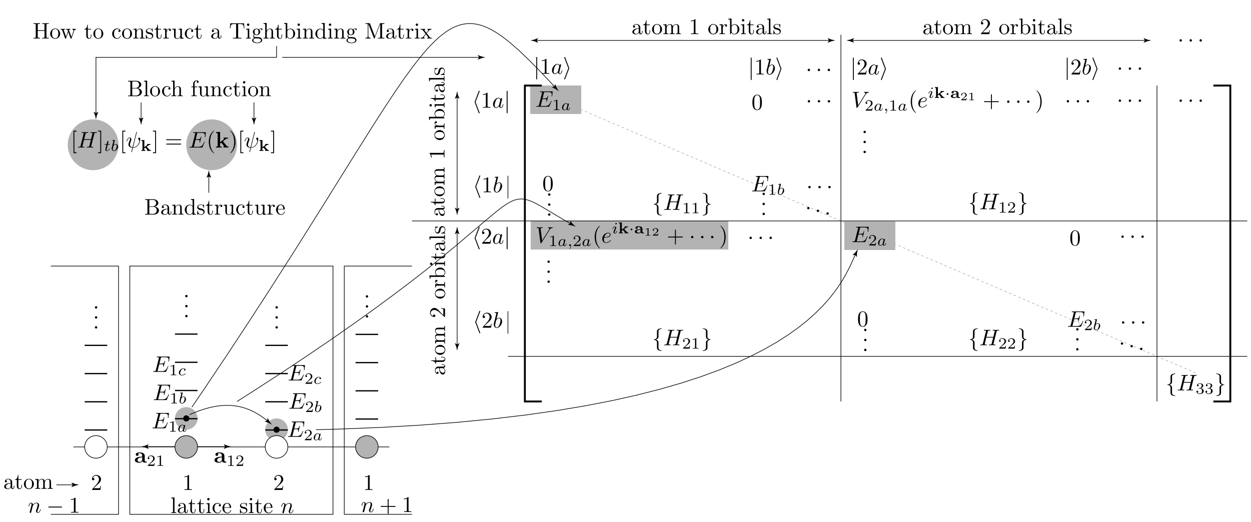

How to construct a Tightbinding Matrix
Bloch function
$[H]_{tb}[\psi_{\mathbf{k}}] = E(\mathbf{k})[\psi_{\mathbf{k}}]$
Bandstructure
atom 1 orbitals
atom 2 orbitals
$|1a\rangle$
$|1b\rangle$
$|2a\rangle$
$|2b\rangle$
$\langle 1a|$
$\langle 1b|$
$\langle 2a|$
$\langle 2b|$
E_{1a}
$V_{2a,1a}(e^{i\mathbf{k}\cdot\mathbf{a}_{21}} + \cdots)$
E_{1b}
$\{H_{11}\}$
$\{H_{12}\}$
$V_{1a,2a}(e^{i\mathbf{k}\cdot\mathbf{a}_{12}} + \cdots)$
E_{2a}
E_{2b}
$\{H_{21}\}$
$\{H_{22}\}$
$\{H_{33}\}$
E_{1c}
E_{1b}
E_{1a}
E_{2c}
E_{2b}
E_{2a}
$\mathbf{a}_{21}$
$\mathbf{a}_{12}$
atom → 2
$n-1$
1 2
lattice site n
1
$n+1$

Fig. 11.9 Anatomy of a tightbinding bandstructure matrix, showing the origin of each matrix element.

For example, in 3D FCC silicon, the basis has two Si atoms, and (1 s + 3 p =) 4 valence orbitals of each Si atom participate in chemical bonding. We can write down Equation 11.10 for each band, but that will be an uninspiring mess of multiple summations and overlap integrals. This mess cleans up instantly if we realize that every tightbinding bandstructure problem really is a matrix problem in disguise! In Fig. 11.9 is shown the anatomy of a tightbinding matrix that rules them all. Let's discuss the construction of the matrix before unleashing its power.

The Bloch wavefunction of Equation 11.10 is written as a column matrix $[\psi_{\mathbf{k}}] = [c_{1a}, c_{1b}, ..., c_{2a}, c_{2b}, ...]^T$, and the tight-binding Hamiltonian equation written in the form of a matrix now reads

$$[H]_{tb}[\psi_{\mathbf{k}}] = E(\mathbf{k})[\psi_{\mathbf{k}}]. \tag{11.11}$$

Upon solving this matrix equation, we obtain two distinct parts of the solution:

Eigenvalues: For any choice of $\mathbf{k}$, there are exactly $N_{\text{atom}} \times N_{\text{orbital}}$ number of eigenvalues $E_n(\mathbf{k})$. As we vary $\mathbf{k}$, the eigenvalues vary, developing into $N_{\text{atom}} \times N_{\text{orbital}}$ bands, which is the size of a side of the $[H]_{tb}$ square matrix. We can label the bands from $n = 1, 2, ..., N_{\text{atom}} \times N_{\text{orbital}}$. For example, for silicon with $N_{\text{atom}} = 2$ atoms/lattice site, and $N_{\text{orbital}} = 4$ per atom, there are $2 \times 4 = 8$ bands. Of course this depends on how many orbitals we decide to include: a more popular and accurate tight-binding model for silicon uses $N_{\text{orbital}} = 5$ orbitals/atom, leading to 10 bands. Similarly, for GaN, there are $N_{\text{orbital}} = 4$ atoms/basis, and $N_{\text{orbital}} = 4$ orbitals/atom, which leads to a 16×16 matrix giving us 16 bands. These eigenvalues are directly measured experimentally by tracking optical transitions, just as in atoms. We will explicitly see these in the following sections.

Eigenfunctions: The second part of the solution is the eigenfunction itself. Since we know the atomic orbital wavefunctions ϕ_n of each atomic state we started with, all we need to know to find the complete Bloch wavefunction of Equation 11.10 are the coefficients $c_{1a}, c_{1b}, ..., c_{2a}, c_{2b}, ...$ When the matrix equation gives us the eigenvalues, corresponding to each eigenvalue $E_n(\mathbf{k})$, the solution also gives its eigenfunction $[\psi_{\mathbf{k}}] = [c_{1a}, c_{1b}, ..., c_{2a}, c_{2b}, ...]^T$. The required coefficients are sitting neatly in this column vector, which we can now use to construct the entire Bloch eigenfunction of Equation 11.10. We can plot it in real space (or **k**-space) to visualize the wavefunctions – they should qualitatively look like Fig. 11.8 in spirit. Plotting $|\psi_{\mathbf{k}}|^2$ will give us the probability distribution of the electron states in real space. If these states are occupied with probability $f(k)$, their corresponding electron distribution in space is $|\psi_k(r)|^2 f(k)$. This physical electron distribution can be experimentally measured by X-ray diffraction techniques. The combination of the eigenvalues and the eigenfunctions can explain the electronic and photonic properties of the crystal. For example, the optical reflectivity and transmission, electronic transport properties, Berry-phases, and topological effects can now all be explained, and even *predicted* from the combination of $E(\mathbf{k})$ and $\psi_{\mathbf{k}}(\mathbf{r})$.

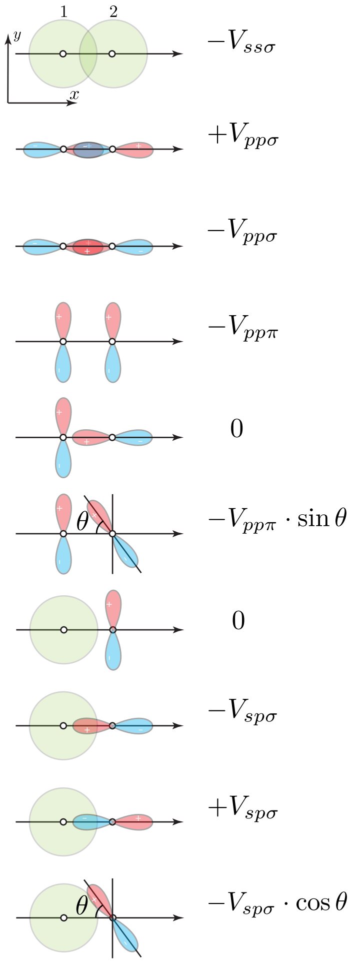


Fig. 11.10 Geometric dependence of tight-binding matrix elements.

The construction of each matrix element in this equation is shown in Fig. 11.9, which connects the matrix elements to the atomic orbitals and the hopping terms and their phases. The diagonal elements are the atomic orbital eigenvalues before the formation of the chemical bonds. If the atoms were very far apart, the matrix is diagonal, with each block containing the eigenvalues of the orbitals. The off-diagonal terms are all due to hopping of electrons from an orbital of one atomic site to an orbital of another atomic site. Because the on-site orbitals are othogonal, these off-diagonal matrix elements are zero. As an example, for electrons in orbital a of atom 1 to hop to orbital a of atom 2, the matrix element is $e^{i\mathbf{k}\cdot\mathbf{a}_{12}} V_{1a2a}$, where $\mathbf{a}_{12} = \mathbf{R}_1 - \mathbf{R}_2$, and $V_{1a2a} = \int \phi_1(r - R_1) W(r) \phi_1(r - R_2)$ is the overlap integral in energy units. If atom 1 has four equivalent atom 2 nearest neighbors labeled 2, 3, 4, and 5, the matrix element will be $(e^{i\mathbf{k}\cdot\mathbf{a}_{12}} + e^{i\mathbf{k}\cdot\mathbf{a}_{13}} + e^{i\mathbf{k}\cdot\mathbf{a}_{14}} + e^{i\mathbf{k}\cdot\mathbf{a}_{15}}) V_{1a2a}$. Note that the nearest neighbor distances $\mathbf{a}_{12}$ enter as vectors, leading to interference of electron waves depending on the direction of their hopping. For example, in a monoatomic linear 1D chain of interatomic distance a, there are two nearest neighbors at $x = (n+1)a$ and $x = (n-1)a$, so the sum is $(e^{ika} + e^{-ika}) V_{1a2a} = 2\cos(ka) \cdot V_{1a2a}$. So a block like $[H]_{21}$ in the tight binding matrix is a Hermitian conjugate of block $[H]_{12}$, $[H]_{21} = ([H]_{12}^T)^\star$, which ensures that the entire matrix is Hermitian, and the matrix thus has real eigenvalues.

The final piece of the puzzle is to reason out what the hopping integrals such as $V_{1a2a}, V_{1a2b}, ...$ are. The matrix elements are the overlap integrals between the orbital wavefunctions between sites. They are all of the form $V_{ij} \sim \eta \frac{\hbar^2}{m_e d^2}$, where η is a factor of the order of unity, $\hbar$ is the reduced Planck's constant, m_e the free–electron energy, and d

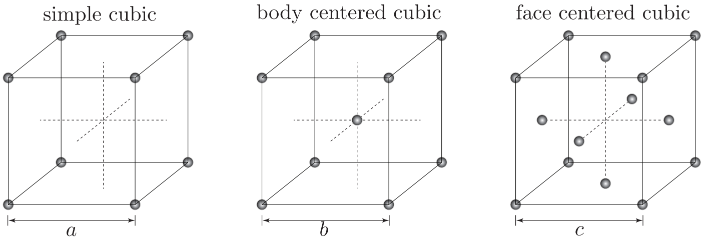


Fig. 11.11 Three crystal structures: simple cubic (SC), body-centered-cubic (BCC), and face-centered cubic (FCC). The side of the cubes are of lengths a, b, and c respectively.

is the distance between the two atoms. They are sometimes called the Slater-Koster matrix elements after an influential work of Slater (Fig. 11.1) and Koster in 1954. The form should look familiar: it is similar to a particle in a box energy, of box size d. There are several competing ways to find these matrix elements – be it between s ↔ s, s ↔ p, or s ↔ d orbital hoppings. Some depend on heuristic arguments, others evaluate them from first principles density functional theory. But the most accurate are the ones that take the help of experimental data. Signatures in the optical spectra due to transitions between energy eigenvalues at specified **k** points provide the information necessary to "measure" the matrix elements. The entire bandstructure can today be measured using Angle-Resolved Photoelectron Spectroscopy (ARPES), which further helps identifying the matrix elements accurately. Fig. 11.10 shows how the matrix elements depend on the overlap geometries of s- and p-orbitals. Note the difference between the $pp\sigma$ and the $pp\pi$ bonds: because of stronger overlap in the bonding $pp\sigma$ case, $V_{pp\pi} < V_{pp\sigma}$ in energy. The overlap integrals $\langle s_1|W|p_{y2}\rangle$ and $\langle p_{y1}|W|p_{x2}\rangle$ are both zero by symmetry. If the p-orbital axis is tilted at an angle θ as shown, the non-zero projection sine or cosine appears in the matrix element. For example, in the last s–p overlap, the p-orbital projections along the line joining the nuclei and perpendicular to it, $|p\rangle \rightarrow |p_x\rangle \cos\theta + |p_y\rangle \sin\theta$ lead to the matrix element $\langle s|W|p\rangle = -V_{sp\sigma} \cdot \cos\theta + 0 \cdot \sin\theta = -V_{sp\sigma} \cdot \cos\theta$.

Bandstructures of simple cubic, BCC, and FCC crystals: Let us compare the tight-binding electronic bands of crystals of three different lattices: the simple cubic (SC), the body-centered-cubic (BCC), and the face-centered cubic (FCC). The three structures are shown in Fig. 11.11. The simplified bandstructures for a single band of each of the three lattices is first obtained using the tight-binding method. Then we will compare their electronic properties. Assume that each sphere shown above is an atom, and they are all the same atoms. Table 11.1 shows the physical properties of the three lattice types. The density of lattice points (# of lattice points/volume) for each crystal structure shows that the density of close-packing increases from simple cubic to BCC to FCC. They can be compared to each other directly if $a = b = c$. Let us consider the tightbinding bandstructure with only an $|s\rangle$ orbital attached to each atom. If the on-site orbital energy is E_s and the nearest neighbor hopping integral is V_{ss} for all three crystals, the resulting

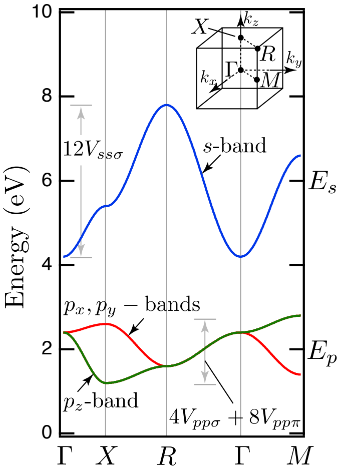


Fig. 11.12 Tight-binding s-bands and p-bands for the 3D simple cubic lattice structure. The reciprocal lattice points are shown. The on-site energies are $E_s = 6$ eV, $E_p = 2$ eV, and hopping energies are $V_{ss\sigma} = 0.3$ eV, $V_{pp\sigma} = 0.3$ eV,, and $V_{pp\pi} = 0.05$ eV. Hopping between $s \leftrightarrow p$ orbitals is neglected.

Table 11.1 Properties of simple cubic, BCC, and FCC lattices.

Property	SC	BCC	FCC
Lattice Density	$\frac{1}{a^3}$	$\frac{2}{b^3}$	$\frac{4}{c^3}$
Lattice Vectors $\mathbf{R}_n$	$a(\pm1,0,0)...$	$\frac{b}{2}(\pm1,\pm1,\pm1)$	$\frac{c}{2}(\pm1,\pm1,0)...$
Nearest Neighbors	6	8	12
s-Band $E_s(k)$	$E_{sc}(k) = E_s - 2V_{ss}[\cos(k_x a) + \cos(k_y a) + \cos(k_z a)]$	$E_{bcc}(k) = E_s - 8V_{ss}[\cos(\frac{k_x b}{2}) \cdot \cos(\frac{k_y b}{2}) \cdot \cos(\frac{k_z b}{2})]$	$E_{fcc}(k) = E_s - 4V_{ss}[\cos(\frac{k_x c}{2}) \cdot \cos(\frac{k_y c}{2}) + \cos(\frac{k_y c}{2}) \cdot \cos(\frac{k_z c}{2}) + \cos(\frac{k_z c}{2}) \cdot \cos(\frac{k_x c}{2})]$
$E(\Gamma) = E(0,0,0)$	$E_s - 6V_{ss}$	$E_s - 8V_{ss}$	$E_s - 12V_{ss}$
Band extrema	$E_s \pm 6V_{ss}$	$E_s \pm 8V_{ss}$	$E_s \pm 12V_{ss}$
Bandwidth	$12V_{ss}$	$16V_{ss}$	$24V_{ss}$
Band-Edge $m^\star$	$\frac{\hbar^2}{2V_{ss}a^2}$	$\frac{\hbar^2}{2V_{ss}b^2}$	$\frac{\hbar^2}{2V_{ss}c^2}$

bandstructures will be each of the form

$$E(k) = E_s - V_{ss} \sum_n e^{i\mathbf{k}\cdot\mathbf{R_n}} = E_s - V_{ss}F(\mathbf{k}), \tag{11.12}$$

where $\mathbf{k} = (k_x, k_y, k_z)$, $\mathbf{R_n}$ are the real-space vectors connecting an atom to its n nearest neighbors, and $F(\mathbf{k}) = \sum_n e^{i\mathbf{k}\cdot\mathbf{R_n}}$ is a hopping function that is characteristic of the lattice structure. Identifying the nearest-neighbor lattice points in Table 11.1, I invite you to show that the s-orbital bandstructure of the SC lattice is $E_{sc}(k) = E_s - 2V_{ss}[\cos(k_x a) + \cos(k_y a) + \cos(k_z a)]$, of the BCC lattice is $E_{bcc}(k) = E_s - 8V_{ss}[\cos(\frac{k_x b}{2}) \cdot \cos(\frac{k_y b}{2}) \cdot \cos(\frac{k_z b}{2})]$, and of the FCC lattice is $E_{fcc}(k) = E_s - 4V_{ss}[\cos(\frac{k_x c}{2}) \cdot \cos(\frac{k_y c}{2}) + \cos(\frac{k_y c}{2}) \cdot \cos(\frac{k_z c}{2}) + \cos(\frac{k_z c}{2}) \cdot \cos(\frac{k_x c}{2})]$. The bandstructures, their extrema, bandwidths, and the effective masses at the band extrema (e.g. at $\Gamma = (0,0,0)$) of each type of crystal structure are listed in Table 11.1. Fig. 11.12 shows the bands of the simple cubic lattice with its Brillouin zone. The s-orbital overlaps at each site lead to one band, and the three p-orbitals at each site lead to three distinct p-bands. Note again how at the Γ-point, the s-bands *curve up*, whereas the p-bands *curve down*: now you know exactly why! Though we have not derived the p-bands explicitly here, it is a nice little exercise you can do to test your understanding.

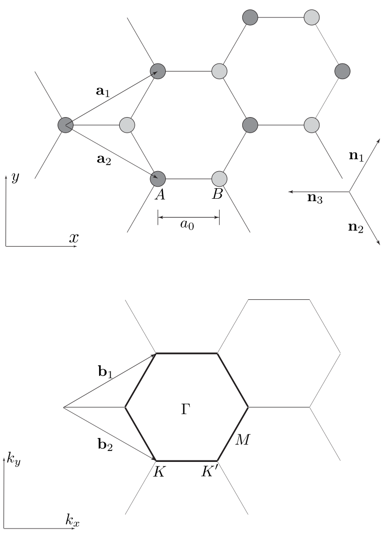


Fig. 11.13 The 2D hexagonal lattice and its Brillouin zone.

11.3 1D, 2D: nanotubes, graphene, BN, MX$_2$

Fig. 11.13 shows the real-space and reciprocal space lattices for a 2D hexagonal crystal. 2D boron nitride and graphene take this crystal structure. The atomic basis has *two* atoms[12] labeled A and B in Fig. 11.13, and each atom has three nearest neighbors.

2D graphene: Let us first consider graphene, for which the basis atoms are both C. Each C atom has four valence electrons in 2s and

[12] Why is the basis of the graphene crystal two C atoms and not one? This is because the hexagonal atom sites do *not* form a 2D Bravais lattice, but the alternate lattice points fit the hexagonal Bravais lattice. In even simpler language: It is not possible to fill a 2D surface by merely *translating* an equilateral triangle, it is necessary to *rotate* the triangle too, which is not allowed. Thus the smallest translating unit area is the parallelogram formed from two triangles.

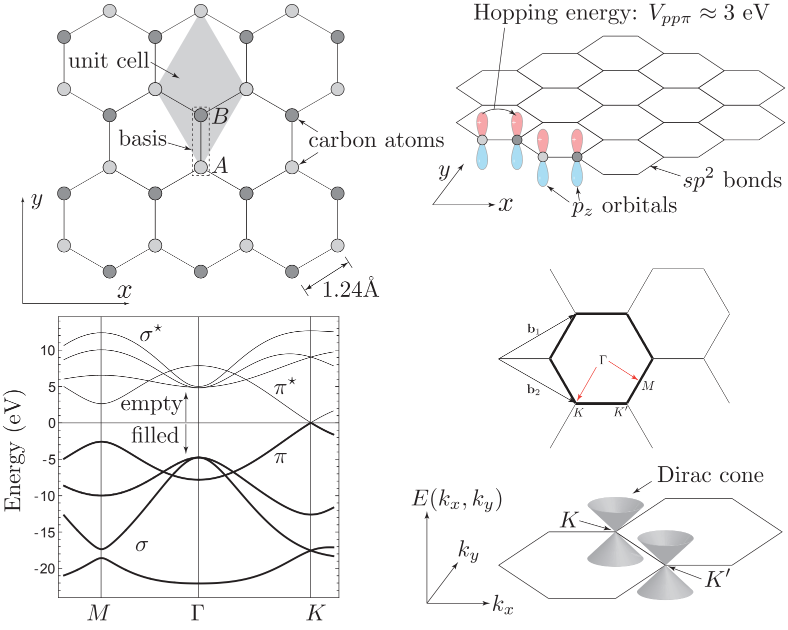


Fig. 11.14 Bandstructure of 2D graphene. The two atoms are B and N for BN, and C and C for graphene. The whole 2D crystal may be generated by repeating this two-atom basis, by translating by the two primitive lattice vectors $\mathbf{a_1} = a_0(\frac{3}{2}, \frac{\sqrt{3}}{2})$ and $\mathbf{a_2} = a_0(\frac{3}{2}, -\frac{\sqrt{3}}{2})$, where a_0 is the distance between the A and B atoms. The three vectors from any atom to the three nearest neighbors are $\mathbf{n}_1 = \frac{a_0}{2}(1, \sqrt{3})$, $\mathbf{n}_2 = \frac{a_0}{2}(1, -\sqrt{3})$, and $\mathbf{n}_3 = -a_0(1, 0)$.

2p orbitals, and the nearest neighbor distance is $a_0 = 1.24$ Å. Since it is known from experiments that the crystal structure is hexagonal, the chemical bonding is identified to be sp^2, which results in hexagonal bonding, something also seen in molecules such as benzene (C_6H_6). A sp^2 bond involves three electrons/atom, so there is one electron/C atom left over, and it is in the p_z orbital. Fig. 11.14 shows the lobes of the probability density $|\psi_{p_z}(r)|^2$ of the p_z orbitals sticking out of the 2D graphene plane. The hopping is thus of the $V_{pp\pi}$ kind between the nearest neighbor p_z orbitals from Fig. 11.10.

We can now set up the tight-binding matrix for calculating the bandstructure of 2D graphene. Taking advantage of the fact that the sp^2 bonds in the $x - y$ plane are orthogonal to the p_z orbitals, the tight-binding Hamiltonian is written as:

$$\begin{array}{c} \\ \langle s^A| \\ \langle p_x^A| \\ \langle p_y^A| \\ \langle s^B| \\ \langle p_x^B| \\ \langle p_y^B| \\ \langle p_z^A| \\ \langle p_z^B| \end{array}
\begin{array}{c} \begin{array}{cccccccc} |s^A\rangle & |p_x^A\rangle & |p_y^A\rangle & |s^B\rangle & |p_x^B\rangle & |p_y^B\rangle & |p_z^A\rangle & |p_z^B\rangle \end{array} \\
\left(\begin{array}{cccccccc}
E_s & 0 & 0 & -V_{ss}g_0(\mathbf{k}) & ... & ... & 0 & 0 \\
0 & E_p & 0 & ... & ... & ... & 0 & 0 \\
0 & 0 & E_p & ... & ... & ... & 0 & 0 \\
\text{c.c.} & \text{c.c.} & \text{c.c.} & E_s & 0 & 0 & 0 & 0 \\
\text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & E_p & 0 & 0 & 0 \\
\text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & 0 & E_p & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & E_p & V_{pp\pi}g_0(\mathbf{k}) \\
0 & 0 & 0 & 0 & 0 & 0 & V_{pp\pi}g_0^\star(\mathbf{k}) & E_p
\end{array}\right) \end{array} .$$

The matrix[13] decomposes into a 6×6 one for the sp^2 bonding, and a 2×2 one for the p_z orbitals. The off-diagonal 3×3 sections

[13] Note the structure of the matrix: since there are four valence orbitals/atom, and two atoms/basis, we have a 8×8 matrix. We have chosen to arrange the matrix in a suggestive way, borrowing from our knowledge of the experimental crystal structure. The sp^2 orbitals at sites A and B are arranged first, and the orthogonal p_z orbitals are included in the last two rows and columns.

that represent the sp^2 overlaps are easily written down by drawing the orbitals at the atomic sites, and using Fig. 11.10. Because the matrix is Hermitian, the lower off-diagonal 3×3 section for the sp^2 bonds is the conjugate transpose of the upper. The matrix element $\langle s^A|W|s^B\rangle = -V_{ss}(e^{i\mathbf{k}.\mathbf{n}_1} + e^{i\mathbf{k}.\mathbf{n}_1} + e^{i\mathbf{k}.\mathbf{n}_3}) = -V_{ss}g_0(\mathbf{k})$. Note how if the nearest neighbor hoppings are all positive in sign, the phase term is $g_0(\mathbf{k}) = e^{i\mathbf{k}.\mathbf{n}_1} + e^{i\mathbf{k}.\mathbf{n}_1} + e^{i\mathbf{k}.\mathbf{n}_3}$. Here $\mathbf{k} = (k_x, k_y)$, and the dot product with the three nearest neighbor vectors compose the phase term. The sign of the hopping integral can be positive or negative, depending on the orientation of the bonds – meaning the phase terms are all of the form $g(\mathbf{k}) = (\pm e^{i\mathbf{k}.\mathbf{n}_1} \pm e^{i\mathbf{k}.\mathbf{n}_1} \pm e^{i\mathbf{k}.\mathbf{n}_3})$. Consider it an exercise to write down the eight remaining matrix elements in the upper off-diagonal sp^2 block. This will involve V_{sp}, $V_{pp\sigma}$, and $V_{pp\pi}$ hopping integrals, and you will need to use Fig. 11.10 generously. The known matrix elements for graphene are: $E_s = -17.52$ eV, $E_p = -8.97$ eV, $V_{ss\sigma} = 4.51$ eV, $V_{sp\sigma} = 5.93$ eV, $V_{pp\sigma} = 5.80$ eV, and $V_{pp\pi} = 2.61$ eV. Using these values, the eigenvalues of the tightbinding matrix are calculated and plotted for the $M \leftrightarrow \Gamma \leftrightarrow K$ points in Fig. 11.14. The 8×8 matrix gives eight eigenvalues for each $\mathbf{k}$-point, which join to form eight bands. Since there are four electrons/atom, and two atoms/basis, there are eight electrons/unit cell that need to fill these bands. Since each band can take $2N$ electrons, four of the lowest bands will be filled. The filled states are shown in bold lines, and the empty states are shown in lighter lines. The highest filled state in intrinsic graphene is at the K and K' points of the BZ. There is no bandgap, because the lowest empty state is at the same energy, and at the same K and K' points. The bandstructure of the π-band around these hexagonal vertices of the BZ is *conical*, and well separated from the σ and $\sigma^\star$ bands, by a few eV. The electrical conductivity of a crystal is dominated by the electron states around the Fermi level. So we conclude that, rather remarkably, the sp^2 bonds of graphene hold the atoms together via chemical bonds, but do not directly contribute to the electronic conductivity. The electrical conductivity is dominated by the $p_z - p_z$ π bonds. To investigate electron transport, or low-energy optical properties of graphene, it is not necessary to include the sp^2 σ bonds, and we can focus just on the p_z bands – which means we will be OK investigating just the 2×2 matrix in the bottom right corner!

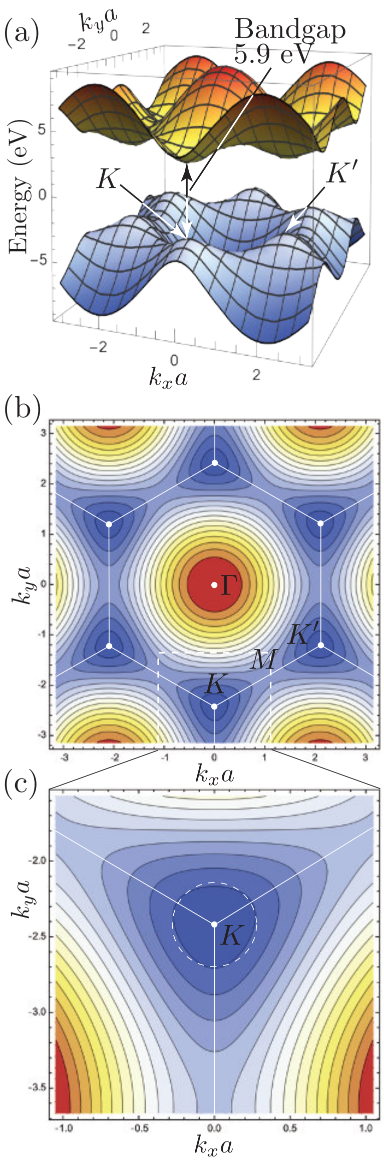


Fig. 11.15 (a) Bandstructure of 2D boron nitride showing energy gaps, (b) constant energy contours of the conduction band, and (c) close-up of a K-valley conduction band minimum showing slight warping from a circular shape.

Let us generalize this 2×2 matrix by allowing on-site orbital energies E_A and E_B, and a hopping energy $-V_{pp\pi}$. The off-diagonal (hopping) term is $g_0(\mathbf{k}) = -V_{pp\pi}(e^{i\mathbf{k}\cdot\mathbf{n}_1} + e^{i\mathbf{k}\cdot\mathbf{n}_2} + e^{i\mathbf{k}\cdot\mathbf{n}_3})$, and the resulting bandstructure are the 2×2 matrix eigenvalues:

$$E_{\pm} = A \pm \sqrt{B^2 + V_{pp\pi}^2[1 + 4\cos(\frac{3}{2}k_x a_0)\cos(\frac{\sqrt{3}}{2}k_y a_0) + 4\cos^2(\frac{\sqrt{3}}{2}k_y a_0)]}, \tag{11.13}$$

where $A = \frac{E_A + E_B}{2}$ and $B = \frac{E_A - E_B}{2}$. For graphene, $E_A = E_B = E_p$, implying $A = E_p$, $B = 0$, and the bandstructure when expanded around the K or K' points for small values of $(q_x, q_y) = (k_x - K_x, k_y - K_y)$ becomes $E_\pm = E_p \pm \hbar v_F \sqrt{q_x^2 + q_y^2}$, which is indeed a cone. This form of bandstructure is popularly known as a Dirac cone because it looks just like the Dirac equation for an electron. For graphene, we can assume $E_A = E_B = 0$ eV, and from experiments we know that $V_{pp\sigma} = 3.0$ eV and $a_0 = 0.15$ nm. Since the bandstructure is conical, the group velocity $\mathbf{v}_g(\mathbf{k}) = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}) = v_F \hat{\mathbf{r}}$ is a *constant* at various $\mathbf{k}$ values $\neq 0$, pointing radially outwards from the K and K' points. In this sense, electrons in 2D graphene behave as light: the energy is *linearly* proportional to the crystal momentum, and one cannot define an effective mass. The electron states are sometimes referred to ”massless Dirac fermions” in the literature. The 1D carbon nanotube is a graphene sheet rolled into a cylinder, so the bandstructure of carbon nanotubes are single k-line cuts of the 2D bandstructure of graphene, and are obtained directly from the graphene bandstructure. If the k-line cute passes through the K or K' points, the nanotube is metallic, and if it misses these points, the nanotube is semiconducting.

2D BN: If we consider the closely related 2D crystal: monolayer BN, in which the two atoms in the basis are B and N. The on-site p_z orbital energies are now different: $E_A = E_{Bp}$ and $E_B = E_{Np}$. Using $E_A = +2.9$ eV, $E_B = -2.9$ eV, $V_{pp\pi} = -2.9$ eV, and $a_0 = 0.15$ nm, we can plot the bandstructure for 2D BN as shown in Fig. 11.15. Note the drastic effect of breaking the symmetry of the basis atoms: this has resulted in the opening of a *large* energy bandgap, of the order $E_g \sim$6 eV! The valence band maxima and the conduction band minima are at the K and K' points, making it a direct bandgap semiconductor. Fig. 11.15 also shows the constant energy contours of the conduction band in the entire BZ, and a zoomed-in version: for very small energies around the K point, these contours are almost circular, but become increasing warped and tend towards triangular for higher energies. Close to the K and K' points, the bandstructure of the conduction and valence bands can be expanded as a parabola for small crystal momentum values, yielding an effective mass $m^\star \approx (2\hbar^2 E_g)/(9a_0 V_{pp\pi}^2) \sim 0.6 m_e$. Since the empty states of the conduction band are separated from the filled states of the valence band by such a large energy gap, 2D BN is an excellent electronic insulator. Since it can be atomically flat, it is attractive as an insulating substrate and atomically thin tunneling barriers. It also provides a periodic potential which breaks the symmetry in the potential between the A and B sites, which is typically referred to as a broken sublattice symmetry potential.

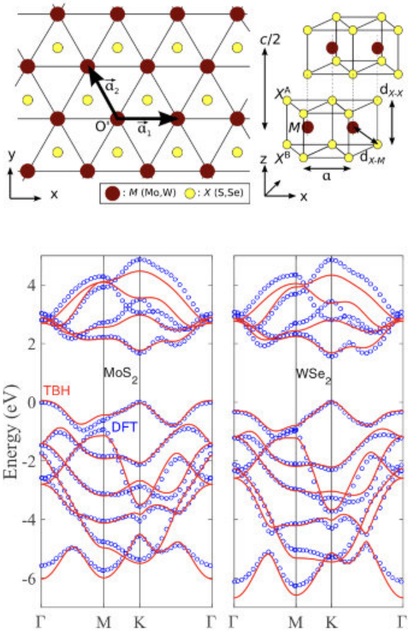


Fig. 11.16 Crystal structure and Electronic Bandstructure of monolayer 2D semiconductors MoS_2 and WSe_2. From Physical Review B, vol. 92, page 205108 (2015).

2D MX_2 crystals: Several other families of 2D semiconductors exist, such as the II-VI InSe, GaSe, black Phosphorus, and the transition metal dichalcogenides. Some of them are of the chemical formula MX_2, where M can be a transition metal or a group IV element, and X is a chalcogen from group VI. Typical examples are $SnSe_2$, MoS_2, and WSe_2. MX_2 crystals have the same hexagonal lattice as graphene and BN, but instead of two atoms, the basis is the triatomic X-M-X combination. The M atom is attached to the hexagonal lattice, and the chalcogen atoms X are above and below, as seen in Fig. 11.16. Like graphene and BN, the layers can stack with weak van der Waals interlayer bonding: an example stacking order called the 2H is shown in Fig. 11.16. The stacking sequence affects the bandstructure: the 2H version is semiconducting, whereas the 1T version is metallic. For transition metal dichalcogenides, the orbitals that participate in chemical bonding are the 5 *d*-orbitals of the transition metal atom M (Mo, W), and the 2×3 *p*-orbitals of the chalcogen atom X (S, Se). The matrix size is thus 11×11. Including hopping to next-nearest neighbors increases the accuracy of the tight-binding bandstructure.

The calculated bandstructure is shown in Fig. 11.16 for MoS_2 and WSe_2, superposed with a density–functional calculation. The density–functional bandstructure was used to extract the tight-binding parameters in this calculation. The 11 orbitals lead to 11 bands. The M atom has in the outermost orbital 6 electrons in the shells $nd^5(n+1)s^1$, and each chalcogen atom has 4 electrons in the shell np^4. The basis therefore has 14 valence electrons. The 7 lowest bands are filled with the 14 electrons, and a bandgap of ~ 2 eV is therefore seen in these monolayer MoS_2 and WSe_2 (Fig. 11.16). Note that the valence band maxima occur both at the K point and the Γ point for MoS_2, but only at the K point for WSe_2. The conduction band minima occur at the K point for both, making them direct bandgap semiconductors in the monolayer form. The bottom of the conduction band at K is primarily formed by the *d*-orbitals of the metal atom M. The top of the valence band at K is primarily formed from the *p*-orbital of the chalcogen atom X. As a result, when multiple layers are stacked, the conduction band minima states are almost unaffected as they are localized on the M atoms far from the X-atom surface. On the other hand, the valence band maximum is strongly perturbed: the valence band edges move up in energy, shrinking the bandgap. In the bulk (multilayer) form, the highest valence band energy is at the Γ-point, and the semiconductor has become indirect.

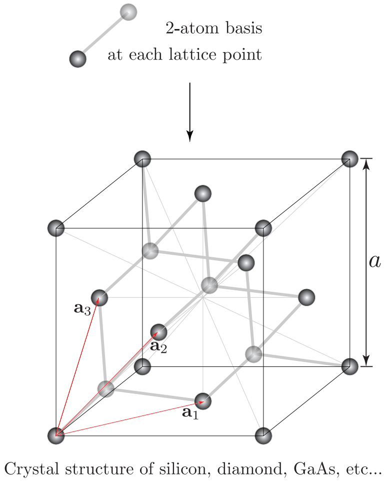


Fig. 11.17 3D FCC crystal structure.

11.4 3D FCC: Si, GaAs

We now extend the tight-binding bandstructure formalism to 3D semiconductors that have a FCC lattice, such as silicon, Ge, GaAs and InP shown in Fig. 11.17[14]. For the family of FCC semiconductors such as silicon and GaAs, the basis has two atoms, but attached to each atom are 4 orbitals: $|s\rangle, |p_x\rangle, |p_y\rangle, |p_z\rangle$, and the chemical bonding is sp^3. The nearest neighbor vectors of the tetrahedral bonds are $\mathbf{n}_1 = \frac{a}{4}(1,1,1)$, $\mathbf{n}_2 = \frac{a}{4}(-1,-1,1)$, $\mathbf{n}_3 = \frac{a}{4}(-1,1,-1)$, $\mathbf{n}_4 = \frac{a}{4}(1,-1,-1)$, where a is the lattice constant ($\neq$ nearest neighbor distance!). Because of various orbital overlaps ss, $sp\sigma$, $pp\sigma$, and $pp\pi$, there are multiple hopping terms. To obtain the energy bandstructure, we should solve a 8×8 matrix numerically. However, a 8×8 formulation is unable to explain why some FCC crystals such as Si, Ge, and GaP are *indirect*

[14] Silicon powers the electronics industry today, and GaAs, InP, and InAs the photonic communications industry – they together are the basis of the new information age. The bandstructure of these semiconductors plays a *central* role in their multifaceted device applications that touch every aspect of our lives today. It is rather remarkable to think that so much can be done with these semiconductors by a knowledge of their bandstructures. All the blood and sweat you have put into the subject till now is finally going to pay off!

bandgap semiconductors. The inclusion of a higher energy orbital, referred to as a $s^\star$ orbital (the nomenclature borrows from the $\sigma^\star$ orbitals of graphene) resolves this issue. The key idea is that the introduction of a higher energy orbital, by perturbation theory, interacts with the eigenvalues of the nearest bands, and due to level repulsion, pushes them down. The strength with which it pushes the lowest conduction band states down is able to explain the experimentally measured bandstructures. Therefore, let us directly deal with the $sp^3s^\star$ tightbinding bandstructure. Since an additional $s^\star$ orbital is introduced to both the basis atom sites of the FCC crystal, the matrix size increases from a 8×8 to a 10×10, and we add just one more hopping term: from the $s^\star$ orbitals to the nearest neighbor p-orbitals, of the form $V_{s^\star p}$. Here is the celebrated 10×10 $sp^3s^\star$ matrix:

$$\begin{array}{c|cccccccccc|}
 & |s^A\rangle & |p_x^A\rangle & |p_y^A\rangle & |p_z^A\rangle & |s^{A\star}\rangle & |s^B\rangle & |p_x^B\rangle & |p_y^B\rangle & |p_z^B\rangle & |s^{B\star}\rangle \\
\langle s^A| & E_s^A & 0 & 0 & 0 & 0 & V_0g_0(\mathbf{k}) & V_1g_1(\mathbf{k}) & V_1g_2(\mathbf{k}) & V_1g_3(\mathbf{k}) & 0 \\
\langle p_x^A| & 0 & E_p^A & 0 & 0 & 0 & -V_1g_1(\mathbf{k}) & V_2g_0(\mathbf{k}) & V_3g_3(\mathbf{k}) & V_3g_2(\mathbf{k}) & -V_5g_1(\mathbf{k}) \\
\langle p_y^A| & 0 & 0 & E_p^A & 0 & 0 & -V_1g_2(\mathbf{k}) & V_3g_3(\mathbf{k}) & V_2g_0(\mathbf{k}) & V_3g_1(\mathbf{k}) & -V_5g_2(\mathbf{k}) \\
\langle p_z^A| & 0 & 0 & 0 & E_p^A & 0 & -V_1g_3(\mathbf{k}) & V_3g_2(\mathbf{k}) & V_3g_1(\mathbf{k}) & V_2g_0(\mathbf{k}) & -V_5g_3(\mathbf{k}) \\
\langle s^{A\star}| & 0 & 0 & 0 & 0 & E_s^{A\star} & 0 & V_4g_1(\mathbf{k}) & V_4g_2(\mathbf{k}) & V_4g_3(\mathbf{k}) & 0 \\
\langle s^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & E_s^B & 0 & 0 & 0 & 0 \\
\langle p_x^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & E_p^B & 0 & 0 & 0 \\
\langle p_y^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & 0 & E_p^B & 0 & 0 \\
\langle p_z^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & 0 & 0 & E_p^B & 0 \\
\langle s^{B\star}| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & 0 & 0 & 0 & E_s^{B\star}
\end{array},$$

where c.c. stands for the complex conjugate (e.g. $H_{71} = H_{17}^\star$).

The constants in this matrix are
$V_0 = V_{ss\sigma}$,
$V_1 = \frac{1}{\sqrt{3}} V_{sp\sigma}$,
$V_2 = \frac{1}{3} V_{pp\sigma} - \frac{2}{3} V_{pp\pi}$,
$V_3 = \frac{1}{3} V_{pp\sigma} + \frac{1}{3} V_{pp\pi}$,
$V_4 = V_{s^\star p}$, and $V_5 = V_{pastc}$.

The nearest neighbor "hopping" phases are

$g_0(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n_1}} + e^{i\mathbf{k}\cdot\mathbf{n_2}} + e^{i\mathbf{k}\cdot\mathbf{n_3}} + e^{i\mathbf{k}\cdot\mathbf{n_4}}$,
$g_1(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n_1}} - e^{i\mathbf{k}\cdot\mathbf{n_2}} - e^{i\mathbf{k}\cdot\mathbf{n_3}} + e^{i\mathbf{k}\cdot\mathbf{n_4}}$,
$g_2(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n_1}} - e^{i\mathbf{k}\cdot\mathbf{n_2}} + e^{i\mathbf{k}\cdot\mathbf{n_3}} - e^{i\mathbf{k}\cdot\mathbf{n_4}}$, and
$g_3(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n_1}} + e^{i\mathbf{k}\cdot\mathbf{n_2}} - e^{i\mathbf{k}\cdot\mathbf{n_3}} - e^{i\mathbf{k}\cdot\mathbf{n_4}}$
for the 4 nearest neighbors of each atom.

Examine the matrix above carefully – you should be able to explain each term based on what we have learnt in this chapter till now. Specifically, all the diagonal terms, and the zeroes are simple – then explain the non-zero off-diagonal terms by invoking symmetries, geometries, and orbital overlaps.

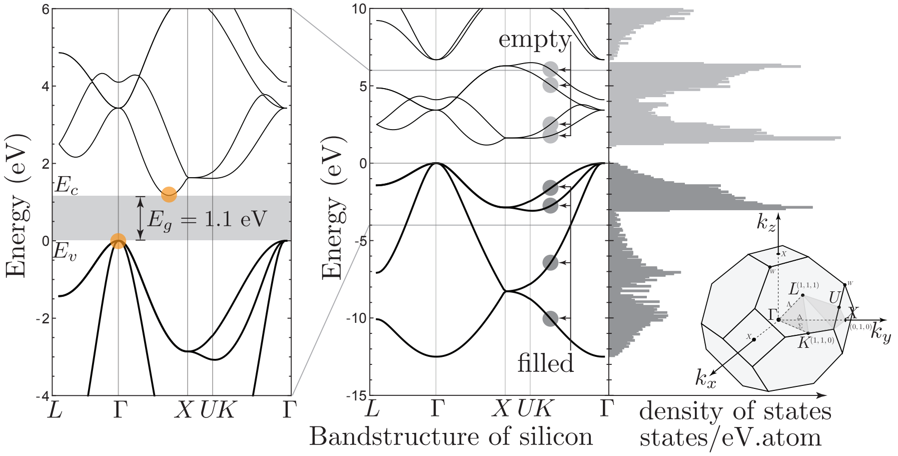


Fig. 11.18 Silicon tight binding bandstructure and density of states using the $sp^3s^\star$ matrix.

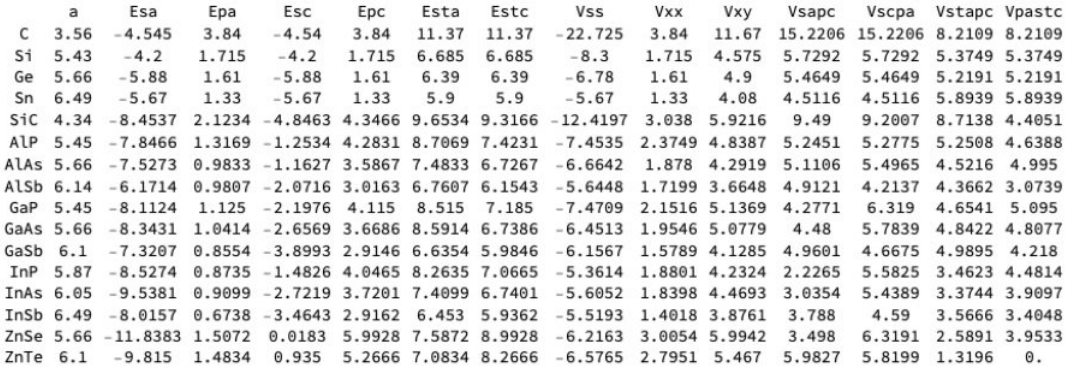
The eigenvalue spectrum of the 10×10 matrix above yields the bandstructure $E(\mathbf{k})$ of silicon, GaAs, and all 3D semiconductors of the diamond-cubic or zinc-blende families. The electron wavevector $\mathbf{k} = (k_x, k_y, k_z)$ is three dimensional: the electron energy depends on the direction it moves. In Chapter 10 we discussed a **k**-space path for FCC crystals $L \leftrightarrow \Gamma \leftrightarrow X \leftrightarrow K \leftrightarrow \Gamma$.

The table in Fig. 11.19 lists all the required constants for the tight-binding matrix. We can now plot the energy bandstructure $E(k)$ for silicon along the chosen **k**-path. It is shown in Fig. 11.18. Again, since we started with 10 orbitals, for each crystal momentum k we get 10 eigenvalues, which join to form 10 bands. Since there are 4 valence electrons/atom, and 2 atoms/basis, there are 8 electrons available for each k. They fill the lowest 4 bands, marked as "filled" in Fig. 11.18. The highest eigenvalue of the valence band is at the Γ point, but the lowest eigenvalue of the empty states is along the $\Gamma - X$ line, near the X-point. The bandgap E_g is the difference of these two eigenvalues, and since they do not have the same k, silicon is an **indirect bandgap semiconductor**. Because light (photons) can cause transitions between electron eigenvalues at the same k, silicon does not emit light efficiently. It can absorb light efficiently, but as the bandstructure shows, the photon energy then must be >3 eV, far larger than its bandgap, in the UV region of the electromagnetic spectrum[15].

The calculated bandstructure directly yields the conduction band edge E_c, the valence band edge E_v, and the bandgap E_g as shown in Fig. 11.18. There are three valence bands degenerate at the Γ-point, which can be seen most clearly along the $K - \Gamma$ direction. They are called the light hole, heavy hole, and the split-off band. We will have more to say about these bands in later chapters.

The minimum of the conduction band occurs in the $\Gamma - X$ direction, meaning there are six such equivalent valleys (Fig. 11.20). So the val-

[15] A peculiar property of the silicon bandstructure is that the difference between the conduction band minimum and the valence band maximum is roughly the same energy over the entire k-path chosen. This bandstructure peculiarity implies photons of this energy can be absorbed very strongly by a large number of electrons, resulting in a large peak in the absorption spectrum. We discuss this further in Chapter 27.



	a	Esa	Epa	Esc	Epc	Esta	Estc	Vss	Vxx	Vxy	Vsapc	Vscpa	Vstapc	Vpastc
C	3.56	-4.545	3.84	-4.54	3.84	11.37	11.37	-22.725	3.84	11.67	15.2206	15.2206	8.2109	8.2109
Si	5.43	-4.2	1.715	-4.2	1.715	6.685	6.685	-8.3	1.715	4.575	5.7292	5.7292	5.3749	5.3749
Ge	5.66	-5.88	1.61	-5.88	1.61	6.39	6.39	-6.78	1.61	4.9	5.4649	5.4649	5.2191	5.2191
Sn	6.49	-5.67	1.33	-5.67	1.33	5.9	5.9	-5.67	1.33	4.08	4.5116	4.5116	5.8939	5.8939
SiC	4.34	-8.4537	2.1234	-4.8463	4.3466	9.6534	9.3166	-12.4197	3.038	5.9216	9.49	9.2007	8.7138	4.4051
AlP	5.45	-7.8466	1.3169	-1.2534	4.2831	8.7069	7.4231	-7.4535	2.3749	4.8387	5.2451	5.2775	5.2508	4.6388
AlAs	5.66	-7.5273	0.9833	-1.1627	3.5867	7.4833	6.7267	-6.6642	1.878	4.2919	5.1106	5.4965	4.5216	4.995
AlSb	6.14	-6.1714	0.9807	-2.0716	3.0163	6.7607	6.1543	-5.6448	1.7199	3.6648	4.9121	4.2137	4.3662	3.0739
GaP	5.45	-8.1124	1.125	-2.1976	4.115	8.515	7.185	-7.4709	2.1516	5.1369	4.2771	6.319	4.6541	5.095
GaAs	5.66	-8.3431	1.0414	-2.6569	3.6686	8.5914	6.7386	-6.4513	1.9546	5.0779	4.48	5.7839	4.8422	4.8077
GaSb	6.1	-7.3207	0.8554	-3.8993	2.9146	6.6354	5.9846	-6.1567	1.5789	4.1285	4.9601	4.6675	4.9895	4.218
InP	5.87	-8.5274	0.8735	-1.4826	4.0465	8.2635	7.0665	-5.3614	1.8801	4.2324	2.2265	5.5825	3.4623	4.4814
InAs	6.05	-9.5381	0.9099	-2.7219	3.7201	7.4099	6.7401	-5.6052	1.8398	4.4693	3.0354	5.4389	3.3744	3.9097
InSb	6.49	-8.0157	0.6738	-3.4643	2.9162	6.453	5.9362	-5.5193	1.4018	3.8761	3.788	4.59	3.5666	3.4048
ZnSe	5.66	-11.8383	1.5072	0.0183	5.9928	7.5872	8.9928	-6.2163	3.0054	5.9942	3.498	6.3191	2.5891	3.9533
ZnTe	6.1	-9.815	1.4834	0.935	5.2666	7.0834	8.2666	-6.5765	2.7951	5.467	5.9827	5.8199	1.3196	0.

Fig. 11.19 Tight binding matrix elements for various FCC semiconductors, from J. Phys. Chem. Solids Vol. 44, pp. 365-378, 1983. a is the lattice constant in Å. The notation connected to the text are: $Esa = E_s^A$ and $Esc = E_s^B$, where the following associations must be made: a =anion, and c =cation. s =s-orbital, p =p-orbital, and st =s* orbital. $V_{ss} = V_0$, $V_{sapc} = V_1$, $V_{xx} = V_2$, $V_{xy} = V_3$, and $V_{stapc} = V_4$.

ley degeneracy of the Si conduction band minimum is $g_v = 6$. The curvature of the band in the $\Gamma - X$ direction is smaller than in the direction perpendicular to it, making the constant energy surface shapes of ellipsoids. This means the electron effective mass at the conduction band minimum is anisotropic – it is heavier in the $\Gamma - X$ direction, and lighter perpendicular to it. We will discuss this anisotropy and its consequences further in Chapter 14. Finally, we can count all the eigenvalues that fall within a specified energy window $(E, E + \Delta E)$ because we now have the entire bandstructure at our disposal. Plotting a histogram of these eigenvalues, we obtain the density of states (DOS) of the semiconductor, as seen in Fig.11.18. Examine the DOS carefully, and relate sharp features to the bands – they follow what you intuitively expect! The DOS is zero in the $E_c - E_v = E_g = 1.1$ eV energy window within the bandgap. The DOS rises as $\sqrt{E}$ as it should for 3D electrons, but only for energies very close to the band edge: in fact that part is not clearly visible in this plot over such a large energy range.

The $sp^3s^\star$ tightbinding method has made us rather powerful: we can now calculate the bandstructures of all the FCC semiconductors of Table 11.19. Fig. 11.21 shows the bandstructures of a range of technologically useful semiconductors. The first panel shows how bandstructures evolve for group-IV elemental FCC crystals as we go down the periodic table. Diamond has a very large indirect bandgap along the $\Gamma - X$ direction. The indirect bandgaps decrease from diamond to Si to Ge. The conduction band minimum of Ge is at the L-point. Note the lowest valence bands are degenerate at the X-point. This is because the same atom occupies both basis sites, and the crystal has inversion symmetry[16]. Finally, Sn does not have a bandgap: it is metallic! It highlights an important trend: bandgaps decrease as we move down the periodic table to heavier elements with more electrons.

The second panel shows the bandstructures of some III-V semicon-

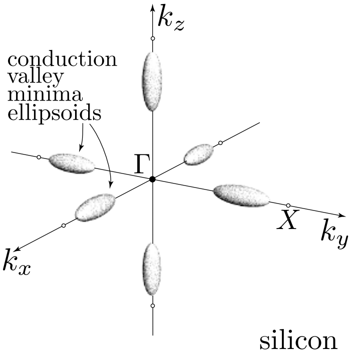


Fig. 11.20 The six equivalent valley minima along the $\Gamma - X$ directions in silicon. These are the "true" silicon valleys!

---

[16] Compare the bottom valence band Dirac-point like crossing at the X-point for Si, Ge with the same eigenvalues for GaAs, InP, and other III-Vs. The fact that there are two different atoms in the basis splits this crossing: can you explain it quantitatively?

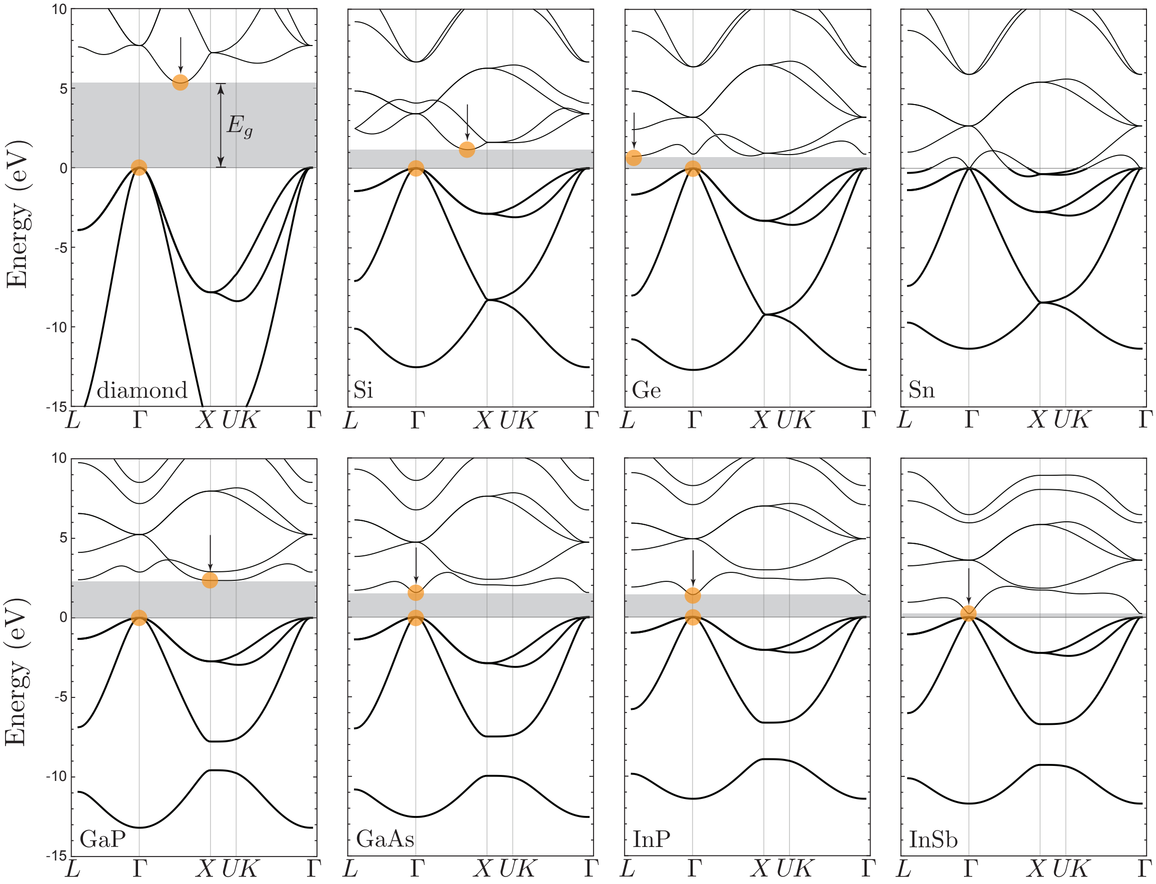


Fig. 11.21 Tight binding bandstructures of various group IV and III-V semiconductors that have the FCC lattice. The conduction band minima are indicated with an arrow to help identify direct vs. indirect bandgap semiconductors. The valence band maxima are always at the Γ-point.

ductors. GaAs, InP, and InSb are direct bandgap semiconductors, since the bottom of the conduction band and the top of the valence band occur at the same crystal momentum state, the Γ point. From our earlier discussions, we can also identify that the bottom of the conduction band must be formed of the *s*-orbitals of the group III metal atom (Ga, In, ...) and the top of the valence band from the *p*-orbitals of the group V atom. An optical transition is now allowed between the CB minimum and the VB maximum with very high efficiency – this is the reason direct-bandgap III-V semiconductors power so many applications in photonics ranging from communications (using lasers and photodetectors), lighting, and imaging. The bandgap of InSb is very small, which means the conduction band curves rather sharply at the minimum, giving it a very small effective mass $m_c^\star$. This feature results in very high electron mobilities, as expected from the expression for electron mobility: $\mu = q\tau/m_c^\star$.

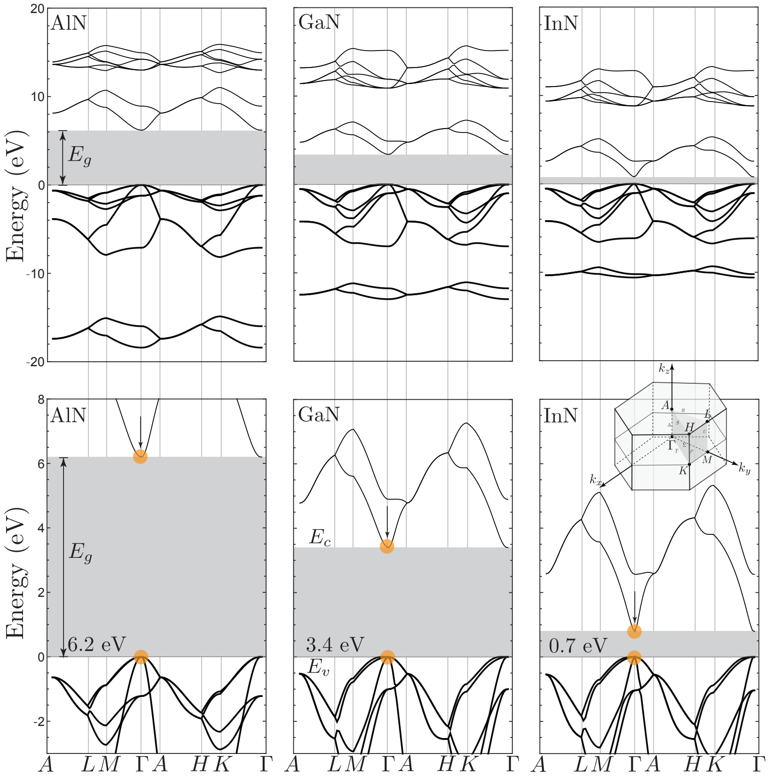


Fig. 11.22 Tight-binding bandstructures of the direct-bandgap nitride semiconductors AlN, GaN, and InN that have the HCP lattice, and a wurtzite crystal structure.

11.5 3D wurtzite: GaN, AlN, ZnO

3D Semiconductors of hexagonal close-packed (HCP) Bravais lattice such as GaN, InN, AlN, and ZnO have become technologically important because of applications in lighting, photonics, and high-speed and high-voltage electronics. These semiconductors crystallize in the wurtzite phase with a 4-atom basis HCP Bravais lattice. For example, the III-V semiconductor GaN has 2 Ga and 2 N atoms in the basis. The chemical bonding is sp^3: Ga atom contributes 3 valence electrons and N atom contributes 5 valence electrons. Thus, there are 16 electrons/basis, meaning the tight-binding matrix is 16×16. In Chapter 10 we outlined the Brillouin zone and the **k**-space path for the HCP lattice. Fig. 11.22 shows the tight-binding bandstructures calculated for AlN, GaN, and InN using parameters in Table 11.2. There are 16 bands of which the lowest 8 are occupied, forming the valence bands. The energy gaps are large for AlN and GaN, as shown in Fig. 11.22. The curvature of the conduction band is sharper than the valence bands. Since the valence band maxima is formed of the 2p orbitals of N, which are deep in energy, the hole effective masses are very large, leading to poor p-type conductivity. But the n-type conductivity of the wide–bandgap nitride semiconductors is comparable to narrower bandgap semiconductors due to the s-orbital origin of their conduction band minima.

Table 11.2 Tight-binding parameters for the calculation of the bandstructures of GaN, AlN, InN, and ZnO. Collected from these references: 1) Physical Review B, vol. 28, page 935 (1983), 2) Physical Review B, vol. 73, page 245327 (2006).

	AlN	GaN	InN	ZnO
E_{sa}	-12.10	-11.01	-6.80	-19.05
E_{pa}	3.58	0.01	0.00	4.14
E_{sc}	-0.10	1.43	-3.02	1.67
E_{pc}	9.42	10.90	8.82	12.37
V_{ss}	-10.74	-5.32	-5.37	-6.04
V_{xx}	5.81	-0.22	0.02	7.16
V_{xy}	8.49	7.14	6.37	10.58
V_{sapc}	8.10	0.63	0.37	4.70
V_{pasc}	9.76	7.28	7.50	8.63

11.6 Tight-binding to design new properties

The bandstructure of semiconductors discussed in this chapter are discussed again in the next two chapters using the $\mathbf{k} \cdot \mathbf{p}$ method, and the Empirical Pseudopotential Method (EPM). The complementary methods help gain experience in the art of bandstructure calculations. Each method provides specific insights that are lacking in the other.

The power of the tightbinding method is the close connection to the actual atoms and the chemical bonds, which we will see will be mostly absent in the other methods. Though tightbinding methods have been, for lack of a better description, "written off" several times in the past, they continue to make comebacks with increased vigor and power. The reason is its root in the chemical bond and chemistry, and no matter how enticing it is to think electrons are mostly free in a solid, they are swimming in a pool of real atoms and nuclei!

The tight-binding method of bandstructure has been used effectively to explore new electronic properties. In the design of artificial superlattices of semiconductors, the tight-binding approach is crucial because it is rather straightforward to make the tight-binding matrix for one semiconductor layer, and connect to another layer by hopping. The concept of *hopping from one atom to the next* is easily extended to *hopping from one semiconductor layer to another*. Because they provide designer electronic, optical, and magnetic properties that are not available in the bulk crystals, such superlattices, multiple quantum wells, or quantum dots are heavily used today in electronic and photonic devices, such as quantum cascade lasers, and will be discussed in the future chapters. The tight-binding method gives us an excellent and intuitive tool to model, design, and predict the layer structures that will lead to targeted physical properties, with a deep connection to the actual atomic and chemical nature of the materials. As an example, we briefly discuss the prediction and discovery of topological insulators and the class of materials that exhibit such phenomena.

Exercise 6.4 discussed the most general 2×2 Hermitian matrix

$$H_2 = \begin{bmatrix} a & b \\ b^\star & c \end{bmatrix} \tag{11.14}$$

which we found can be written as $H_2 = h_0 I_2 + \mathbf{h} \cdot \vec{\sigma}$, where $\mathbf{h} = (h_x, h_y, h_z)$ is a **vector** whose components are three complex numbers, and the three components of $\vec{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are the three Pauli matrices. Now the three complex numbers for a *bandstructure* has the crystal momentum as a *parameter*, meaning in general, $h_x = h_x(k_x, k_y, k_z)$, and similarly for h_y and h_z. For example, for 2D graphene, we can expand the Hamiltonian for small values of (k_x, k_y) around the Dirac point at K to write

$$H_{gr} = E_0 I_2 + \hbar v_F \begin{bmatrix} 0 & k_x - ik_y \\ k_x + ik_y & 0 \end{bmatrix} = E_0 I_2 + \hbar v_F (\underbrace{k_x\sigma_x + k_y\sigma_y + 0\cdot\sigma_z}_{\mathbf{k}\cdot\vec{\sigma}}), \tag{11.15}$$

in which we identify the vector $\mathbf{h} = \hbar v_F(k_x, k_y, 0)$. As we change the values of (k_x, k_y) in the BZ, the tip of the unit vector $\hat{\mathbf{h}} = \mathbf{h}/|\mathbf{h}|$ moves on the equatorial circle. Now the Chern number of a vector field $\mathbf{h}$ is defined as $n = \frac{1}{2\pi}\int_{BZ} d^2k\Omega(k_x, k_y)$, where $\Omega(k_x, k_y)$ is the Berry curvature, defined in Chapter 9 for Bloch states. The Chern number is typically zero, but if it is non-zero, for integrals over the entire Brillouin zone, it is *quantized*, meaning it can only be an integer. It is the number of times the tip of the vector $\hat{\mathbf{h}}$ wraps around the surface of the unit sphere when (k_x, k_y) is varied over the entire Brillouin zone. This connection is able to explain electronic polarization and magnetization in crystals, and if the Chern number is non-zero, predicts the existence of edge states that have no bandgap, but with gapped bulk states: a **topological insulator**.

The tight-binding method for the calculation of bandstructures has thus guided the discovery of a new class of crystals that have topologically non-trivial properties. For example, the 2x2 tight-binding Hamiltonian of graphene has a non-zero Berry phase. It is a point of departure to explore several topologically non-trivial properties of crystals, such as topological insulators, quantum spin-Hall effects, etc. In 1988, Haldane (Fig. 11.23) discovered a way to produce non-zero Chern numbers from the tight-binding picture of graphene. This identification proved to the be precursor to the prediction of topological insulators, again using the 2x2 matrix, or its slight variants. It is safe to say that the tightbinding method of bandstructure still has reserves of secrets to reveal about crystals in the future!

Fig. 11.23 Duncan Haldane discovered that introducing unconventional hopping into the graphene-like tight-binding 2×2 Hamiltonian results in a crystal that in principle can show the quantized Hall effect *without* an external magnetic field. This was the precursor of topological insulators. Haldane was awarded the Nobel Prize in Physics in 2016 with Thouless and Kosterlitz for their contributions to topological phenomena in physics.

11.7 Chapter summary section

In this chapter, we learned:

- How sharp eigenvalues of electrons tightly bound to atoms in a periodic arrays of atoms evolves into electron wavefunction overlaps and spreading of energy eigenvalues into bands to form extended states,
- A deep connection between the Bloch theorem, and the tight-binding molecular picture, between bonding and antibonding orbitals, and the bandstructure of s, p, d orbitals, and its application to several 2D and 3D crystals, and
- To appreciate the power of tight-binding methods to yield insight to the physical properties, and seed ideas for designing new properties.

Further reading

Yu and Cardona's **Fundamentals of Semiconductors** has a wealth of bandstructures to read and explore. Roald Hoffmann's short and sweet book **Solids and Surfaces** is a delightful read, and provides an unmatched physical insight to the tight-binding procedure looking at the problem from an atomic and molecular perspective. Harrison's **Electronic Structure and the Properties of Solids** is a classic reference to learn and appreciate the power of tightbinding models to explain a rich range of physical phenomena.

Exercises

(11.1) **Physics of a 1D semiconductor crystal**

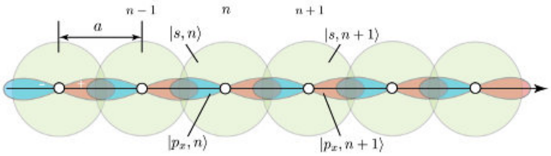


Fig. 11.24 Orbitals and overlaps for a 1D tight-binding lattice.

In this problem, you will calculate the simplified bandstructure of a 1D semiconductor using the tight-binding model. Then you will find the density of states, band-edge effective mass, and solve a quantum transport problem for the semiconductor. Assume a circular ring of atoms with lattice constant a as shown in Fig. 11.24 and N sites, such that $L = Na$. Electrons can occupy two atomic orbitals per lattice site: one $|s\rangle$ and the other $|p_x\rangle$. When the atoms are not in the crystal, the energies of electrons in these orbitals are $H_0|s\rangle = E_S|s\rangle$ and $H_0|p_x\rangle = E_P|p_x\rangle$, such that $E_S > E_P$. Upon forming chemical bonds, the nearest neighbor $|s\rangle$ orbitals hybridize with hopping energy $\langle s, n \pm 1|H_{crystal}|s, n\rangle = -V_{ss\sigma}$, and the nearest neighbor $|p_x\rangle$ orbitals hybridize with hopping energy $\langle p_x, n \pm 1|H_{crystal}|p_x, n\rangle = +V_{pp\sigma}$. Assume in this problem that there is no bonding of $|s, n\rangle$ with $|p_x, n \pm 1\rangle$, meaning $\langle s, n|H_{crystal}|p_x, n \pm 1\rangle = V_{sp\sigma} = 0$.

(a) Show that the two bands formed have bandstructures $E_c(k) = E_S - 2V_{ss\sigma}\cos(ka)$ and $E_v(k) = E_p + 2V_{pp\sigma}\cos(ka)$. We will call the s-band the conduction band, and the p-band the valence band.

(b) Sketch the bandstructure of the crystal in the first Brillouin zone $-\frac{\pi}{a} \leq k \leq +\frac{\pi}{a}$ (remember $E_S > E_P$). Label the extrema points of the bandstructure. Under what conditions will the crystal have no energy gap? Exactly how many electrons can the first BZ hold?

(c) Assume the hopping terms are such that there is a bandgap. What is the bandgap? With how many electrons will you fill the crystal to make it an intrinsic semiconductor?

(d) Using a Taylor expansion of $\cos x \approx 1 - \frac{x^2}{2}$ for $x << 1$, show that the effective mass of the conduction band edge at the Γ-point, i.e. at $k = 0$ is $m_c^\star = \frac{\hbar^2}{2V_{ss\sigma}a^2}$. Relate this effective mass to the total energy bandwidth of the conduction band. Does it make intuitive sense?

(e) By some means, you have managed to populate with $f(k) = 1$ the conduction band states between $-k_L \leq k \leq +k_R$ where $-k_L$ is the highest left-going state, and k_R is the highest right-going state. What is the net quantum current flowing in the crystal?

(f) Show that the exact conduction band DOS is given by $g_c(E) = \frac{g_s}{2\pi a\sqrt{(2V_{ss\sigma})^2-(E-E_S)^2}}$. Make a sketch of this DOS on the side of your bandstructure sketch of part (b) of this problem.

(11.2) **Physics of a 1D crystal**

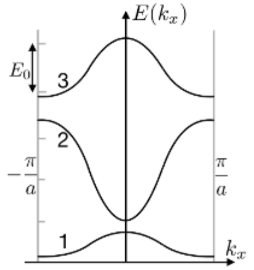


Fig. 11.25 The bandstructure of a 1D crystal showing 3 bands.

(a) Figure 11.25 shows the three lowest three bands of the bandstructure of a 1D crystal of lattice constant a. Copy the figure, and on its right, sketch the corresponding density of states (DOS) of each band. Indicate the differences if any between the DOS of the three bands. **Remember it is a 1D crystal**. Leave some space to the right of this sketch for part (e).

(b) On your copied bandstructure of Fig. 11.25, indicate the state(s) of the highest group velocity. It is given that the lattice constant is $a = 0.5$ nm, and the energy steps are $E_0 = 1$ eV. Give an approximate expression for the highest group velocity, and find its numerical value in cm/s.

(c) You are required to place the Fermi level such that you achieve a n-type semiconductor with a small electron effective mass. Which band will you choose, and where will you place the Fermi level? Roughly how many electrons per unit cell will do the job?

(d) If the Fermi level is near the top of band 2, say at $E_F \approx 3E_0$, and I want to *reduce* the electrical conductivity of the crystal, should I move the E_F up or down? If I have to do this by Coulomb repulsion from a gate, should I apply a positive or a negative voltage on the gate?

(e) If the Fermi level is in the middle of band 2, say $E_F \approx 2E_0$, find the quantum current in the absence of scattering at $T = 0$ K that will flow in the 1D crystal if I apply a voltage of $V = 0.1$ V across Ohmic contacts to its ends. To your sketch of part (a), add a qualitative sketch of the *net current* in the horizontal axis, as a function of the energy in the vertical axis.

(11.3) **Energy bandstructure of 2D graphene and BN**

In this problem, you will quantitatively calculate the electronic bandstructure for 2D graphene and 2D boron nitride as "elemental" and "compound" semiconductors respectively.

(a) Show that if the energies allowed for the electron in individual atoms A and B are E_A and E_B, and the lowering of energy because of hopping of the electron between the atoms between the atoms is U, the energies allowed for the electron after the formation of a chemical bond are $E_\pm = \frac{E_A+E_B}{2} \pm \sqrt{(\frac{E_A-E_B}{2})^2+U^2}$. Because each allowed electron state can hold *two* electrons of opposite spins, what is the *lowering* of energy due to the formation of the bond?

(b) Now consider 2D boron nitride (or graphene), whose atomic basis has *two* atoms: B and N for BN (and C and C for graphene). The whole 2D crystal may be generated by repeating this two-atom basis, by translating by the two primitive lattice vectors $\mathbf{a_1} = a(\frac{3}{2}, \frac{\sqrt{3}}{2})$ and $\mathbf{a_2} = a(\frac{3}{2}, -\frac{\sqrt{3}}{2})$, where a is the distance between the A and B atoms. Sketch the basis, the lattice, and these vectors, and show that the three vectors from any atom to the three nearest neighbors are $\mathbf{n}_1 = \frac{a}{2}(1, \sqrt{3})$, $\mathbf{n}_2 = \frac{a}{2}(1, -\sqrt{3})$, and $\mathbf{n}_3 = -a(1, 0)$. We will look at only the p_z orbital bands. For on-site orbital energies E_A and E_B and a hopping energy t, show that the off-diagonal (hopping) term is $h(\mathbf{k}) = -t(e^{i\mathbf{k}\cdot\mathbf{n_1}} + e^{i\mathbf{k}\cdot\mathbf{n_2}} + e^{i\mathbf{k}\cdot\mathbf{n_3}})$, and the resulting bandstructure is given by Equation 11.13.

(c) Using the following parameters, make 2D plots of the bandstructures of BN and graphene similar to Figure 11.15. For BN: $E_A = +2.9$ eV, $E_B = -2.9$ eV, $t = -2.9$ eV, and $a = 0.15$ nm. For graphene, $E_A = E_B = 0$ eV, $t = 3.0$ eV, $a = 0.15$ nm. Describe the differences and similarities between them. Specifically, find the band-edge *effective masses* and *bandgap* for BN, and the *Fermi velocity* characterizing the slope of the $E - k$ Dirac cone of graphene. Relate these analytically to the known parameters.

(11.4) **2D boron nitride**

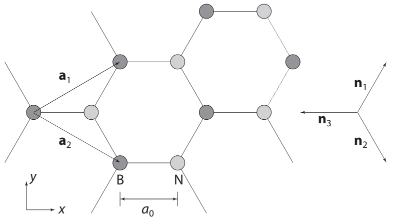


Fig. 11.26 The monolayer boron nitride 2D crystal.

Fig. 11.26 shows a monolayer 2D crystal hexagonal boron nitride. the primitive lattice vectors are $\mathbf{a_1}, \mathbf{a_2}$, the distance between the B and N atoms is a_0, and three related vectors $\mathbf{n_1}, \mathbf{n_2}, \mathbf{n_3}$ are shown to assist you in this problem.

(a) Write $\mathbf{a_1}, \mathbf{a_2}$ in terms of a_0, and $\mathbf{n_1}, \mathbf{n_2}, \mathbf{n_3}$ also in terms of a_0. Find the reciprocal lattice vectors, and sketch the first Brillouin zone.

(b) *Qualitatively* outline a process to find the tight-binding bandstructure of BN.

(c) Of most interest for electronic and photonic properties are the bands formed by the *out-of-plane* p_z orbitals attached to each atom in the crystal. Sketch the orbitals and put signs on the lobes.

(d) In graphene, these orbitals had same energies because all atoms were carbon. Let E_B be the energy of the p_z orbitals on the B sites, and E_N the energy of the p_z orbitals on the N sites. Let $V_{pp\pi}$ be the relevant overlap integral. Write down the tight-binding Hamiltonian for the bands formed by these p_z orbitals.

(e) Find the energy bandgaps at the Γ- point ($k_x = 0, k_y = 0$), and at the *vertices* of the first BZ edge you sketched in part (a) of this problem.

(11.5) **The $sp^3s^\star$ tight-binding bandstructure of silicon, GaAs, and various FCC semiconductors**
In this problem we discuss how to solve a problem of the simple sp^3 8×8 tight-binding model of bandstructure (Exercise 11.6), that is unable to capture the experimental fact that Si and Ge are indirect bandgap semiconductors. Since the exact nature of bandstructure determines all electronic and photonic properties of the semiconductor, it is important to get it right, and it should be a memorable event when you get your own model to calculate the bandstructures.

In their 1983 paper [J. Phys. Chem. Solids, vol 44, pp. 365-378, 1983], Vogl, Hjalmarson and Dow introduced an additional higher energy orbital called $s^\star$ above the conduction band. The major idea is that this level pushes the conduction bands away from the Γ point down by level repulsion just enough to explain why Si and Ge are indirect bandgap whereas GaAs is direct bandgap. So now the sp^3 becomes $sp^3s^\star$, and the 8×8 tightbinding Hamiltonian becomes a 10×10 Hamiltonian.

Read the paper, and reproduce the bandstructures of Si, Ge, and GaAs using your own code. Everything you need is provided in the paper, and also discussed in this chapter. Having done this, you will a very powerful tool at your disposal because with it you can calculate the bandstructure of a very wide range of FCC semiconductors including InSb, diamond as in Fig. 11.21!

(11.6) **Energy bandstructure of silicon and GaAs**
Silicon powers the electronics industry today, and GaAs the photonic industry – they together are the basis of the new information age. You are solving the quantum mechanical problem of both crystals! You will be able to get plots like the ones shown in Fig. 11.21. Because the bandstructure is the allowed electron eigenvalues in the crystal, getting it means you have (almost) completely solved the full quantum mechanical problem of the two semiconductors. Since the bandstructure determines all the electronic and photonic properties of the semiconductor, it should be a memorable event when you get it to work!

Because for graphene and BN we looked at a single $|p_z\rangle$ orbital and a two-atom basis, the matrix was 2×2. For silicon and GaAs, the basis has two atoms too, but attached to each atom will be four orbitals: $|s\rangle, |p_x\rangle, |p_y\rangle, |p_z\rangle$, and the chemical bonding is sp^3. Thus, to get the energy bandstructure we have to solve a 8×8 matrix numerically. Note that unlike in Section 11.4, where we discussed the 10×10 $sp^3s^\star$ bandstructure of Silicon, in this problem we will solve the smaller 8×8 bandstructure, without the two $s^\star$ orbitals.

(a) Sketch the crystal structure, the atomic basis, and the nearest neighbor vectors $\mathbf{n}_1 = \frac{a}{4}(1,1,1)$, $\mathbf{n}_2 = \frac{a}{4}(-1,-1,1)$, $\mathbf{n}_3 = \frac{a}{4}(-1,1,-1)$, $\mathbf{n}_4 = \frac{a}{4}(1,-1,-1)$, where a is the lattice constant ($\neq$ nearest neighbor distance!). Sketch the orbitals centered at each site and identify the overlaps, and the angles between them.

(b) Because of various orbital overlaps ss, $sp\sigma$, $pp\sigma$ and $pp\pi$, there are multiple hopping terms. Examine the matrix carefully and explain its structure. Specifically, explain all the diagonal terms, and the zeroes. Then explain the non-zero off-diagonal terms by invoking symmetries, geometries, and orbital overlaps.

(c) Set up the matrix below in Mathematica, Matlab, Python (or any tool you love!) to plot the energy bandstructure $E(k)$ for k from $\mathbf{L} \leftrightarrow \mathbf{\Gamma} \leftrightarrow \mathbf{X} \leftrightarrow \mathbf{K} \leftrightarrow \mathbf{\Gamma}$. For silicon, $E_s^A = E_s^B = -13.55$ eV, $E_p^A = E_p^B = -6.52$ eV, $V_{ss\sigma} = +2.03$ eV, $V_{sp\sigma} = +2.55$ eV, $V_{pp\sigma} = +4.55$ eV, $V_{pp\pi} = +1.09$ eV, and $a = 0.543$ nm. Compare your result to the more accurate $sp^3s^\star$ calculation shown in Fig. 11.21.

$$\begin{array}{c|cccccccc|}
 & |s^A\rangle & |p_x^A\rangle & |p_y^A\rangle & |p_z^A\rangle & |s^B\rangle & |p_x^B\rangle & |p_y^B\rangle & |p_z^B\rangle \\
\langle s^A| & E_s^A & 0 & 0 & 0 & -V_0 g_0(\mathbf{k}) & V_1 g_1(\mathbf{k}) & V_1 g_2(\mathbf{k}) & V_1 g_3(\mathbf{k}) \\
\langle p_x^A| & 0 & E_p^A & 0 & 0 & -V_1 g_1(\mathbf{k}) & V_2 g_0(\mathbf{k}) & V_3 g_3(\mathbf{k}) & V_3 g_2(\mathbf{k}) \\
\langle p_y^A| & 0 & 0 & E_p^A & 0 & -V_1 g_2(\mathbf{k}) & V_3 g_3(\mathbf{k}) & V_2 g_0(\mathbf{k}) & V_3 g_1(\mathbf{k}) \\
\langle p_z^A| & 0 & 0 & 0 & E_p^A & -V_1 g_3(\mathbf{k}) & V_3 g_2(\mathbf{k}) & V_3 g_1(\mathbf{k}) & V_2 g_0(\mathbf{k}) \\
\langle s^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & E_s^B & 0 & 0 & 0 \\
\langle p_x^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & E_p^B & 0 & 0 \\
\langle p_y^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & 0 & E_p^B & 0 \\
\langle p_z^B| & \text{c.c.} & \text{c.c.} & \text{c.c.} & \text{c.c.} & 0 & 0 & 0 & E_p^B
\end{array},$$

where c.c. stands for the complex conjugate (e.g. $H_{51} = H_{15}^\star$). The constants in this matrix are

$V_0 = V_{ss\sigma}$,
$V_1 = \frac{1}{\sqrt{3}} V_{sp\sigma}$,
$V_2 = \frac{1}{3} V_{pp\sigma} - \frac{2}{3} V_{pp\pi}$, and
$V_3 = \frac{1}{3} V_{pp\sigma} + \frac{1}{3} V_{pp\pi}$,

and the nearest neighbor "hopping" phases are

$g_0(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n}_1} + e^{i\mathbf{k}\cdot\mathbf{n}_2} + e^{i\mathbf{k}\cdot\mathbf{n}_3} + e^{i\mathbf{k}\cdot\mathbf{n}_4}$,
$g_1(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n}_1} - e^{i\mathbf{k}\cdot\mathbf{n}_2} - e^{i\mathbf{k}\cdot\mathbf{n}_3} + e^{i\mathbf{k}\cdot\mathbf{n}_4}$,
$g_2(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n}_1} - e^{i\mathbf{k}\cdot\mathbf{n}_2} + e^{i\mathbf{k}\cdot\mathbf{n}_3} - e^{i\mathbf{k}\cdot\mathbf{n}_4}$, and
$g_3(\mathbf{k}) = e^{i\mathbf{k}\cdot\mathbf{n}_1} + e^{i\mathbf{k}\cdot\mathbf{n}_2} - e^{i\mathbf{k}\cdot\mathbf{n}_3} - e^{i\mathbf{k}\cdot\mathbf{n}_4}$

for the 4 nearest neighbors. The eigenvalue spectrum of the 8×8 matrix above yields the bandstructure $E(\mathbf{k})$ of silicon, GaAs, and all 3D semiconductors of the diamond-cubic or zinc-blende families. The electron wavevector $\mathbf{k} = (k_x, k_y, k_z)$ is three dimensional: the electron energy depends on the direction it moves.

(d) Repeat the bandstructure plot for the compound semiconductor GaAs. Use $E_s^{Ga} = -11.37$ eV, $E_s^{As} = -17.33$ eV, $E_p^{Ga} = -4.9$ eV, $E_p^{As} = -7.91$ eV, $V_{ss\sigma} = +1.70$ eV, $V_{sp\sigma} = +2.15$ eV, $V_{pp\sigma} = +3.44$ eV, $V_{pp\pi} = +0.89$ eV, and $a = 0.565$ nm. Compare your result to the more accurate $sp^3s^\star$ calculation shown in Fig. 11.21.

(e) If you have worked your way to this stage, congratulations, your hard work will pay off handsomely. From the 8×8 tight-binding bandstructure matrix of GaAs, show that for the Γ point with $\mathbf{k} = 0$, the eight eigenvalues can be calculated *analytically*. Then prove that for GaAs, the conduction band edge state is at $E_c(\Gamma) = \frac{E_s^{Ga} + E_s^{As}}{2} + \sqrt{(\frac{E_s^{Ga} - E_s^{As}}{2})^2 + (4V_{ss\sigma})^2}$ composed of $|s\rangle$ orbital overlap between the Ga and As atoms, and no $|p\rangle$ orbitals are involved.

(f) Similarly, show that for the states at the top of the valence band at the Γ point, only $|p\rangle$ orbitals are involved with no $|s\rangle$ orbital involvement. Use this property to argue why holes are anisotropic in most compound semiconductors, but electrons are isotropic.

$\mathbf{k}\cdot\mathbf{p}$ Bandstructure

12

12.1 $\mathbf{k}\cdot\mathbf{p}$ theory 251
12.2 Symmetry 254
12.3 Analytical model without spin 256
12.4 Non-parbolicity and sum rules 259
12.5 The Kane model with spin-orbit interaction 260
12.6 Chapter summary section 266
Further reading 266
Exercises 267

In this chapter, we learn the $\mathbf{k}\cdot\mathbf{p}$ method of calculation of semiconductor bandstructures. In addition to being popular in the design of semiconductor electronic and photonic devices, the method provides powerful new insights into the relation of bandstructure to energy bandgaps, conduction and valence band effective masses, symmetries, and degeneracies. Our goals in this chapter are to learn:

- How to view the electron momentum $\hbar\mathbf{k}$ as a perturbation to the situation when $\mathbf{k} = 0$ and map the bandstructure problem to the perturbation theory result of quantum mechanics,
- Directly apply the results of perturbation theory to obtain the $\mathbf{k}\cdot\mathbf{p}$ bandstructure $E(k)$ for semiconductors,
- How symmetry of states and momentum matrix elements between the valence and conduction bands determine the curvature and effective masses of the bands, and reveal important relations between bandgaps and effective masses, and
- The effect of spin-orbit interaction on the semiconductor bandstructure.

12.1 $\mathbf{k}\cdot\mathbf{p}$ theory

The starting point for the $\mathbf{k}\cdot\mathbf{p}$ bandstructure theory is the Schrödinger equation form of Equation 9.16 from Chapter 9, rewritten here as[1]

$$[\underbrace{-\frac{\hbar^2}{2m_e}\nabla^2 + V}_{\hat{H}^0} + \underbrace{\frac{\hbar}{m_e}\mathbf{k}\cdot\hat{\mathbf{p}} + \frac{\hbar^2k^2}{2m_e}}_{\hat{W}}]u_{\mathbf{k}} = E(\mathbf{k})u_{\mathbf{k}}. \tag{12.1}$$

The symmetries in real space carry over to degeneracies of the allowed energy eigenvalues $E_n(\mathbf{k})$ where n indicates the band index and $\mathbf{k}$ the reduced wavevector in the first Brillouin zone. The Bloch lattice function $u_{\mathbf{k}} = u(\mathbf{k},\mathbf{r})$ at $\mathbf{k} = 0$ inherits *all* the symmetries related to translation *and* rotations inherent in the crystal potential. Since at a *fixed* $\mathbf{k}$, the set of functions $u_n(\mathbf{k},\mathbf{r})$ for various n form a *complete set*, by the expansion principle, if we know the solutions at any *one* $\mathbf{k}$, we can obtain the solutions at *all* $\mathbf{k}$. By splitting the effective Hamiltonian into an "exactly solved" portion $\hat{H}^0$ and a $\mathbf{k}$-dependent part $\hat{W}$, if the solutions $E_n(0)$ and $u_n(0,\mathbf{r})$ at $\mathbf{k} = 0$ are known, then those at non-zero $\mathbf{k}$ can be obtained by perturbation theory.

---

[1] Recall from Chapter 9 that this form is obtained by directly using the Bloch wavefunction $\psi_{\mathbf{k}} = e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}$ in the Schrödinger equation. Therefore this form is sometimes referred to as the **Schrödinger–Bloch equation**. Because the Hamiltonian $\hat{H}^0 + \hat{W}$ is Hermitian, the eigenvalues $E(\mathbf{k})$ are real, and the corresponding eigenfunctions $u_{\mathbf{k}}$ of this form are orthogonal. This $\mathbf{k}\cdot\mathbf{p}$ form of the effective Hamiltonian for crystals and symmetry considerations were first identified by Bouckaert, Smoluchowski, and Wigner in Physical Review 50, 58 (1936). Bardeen applied it to calculate the electron effective mass in metallic sodium and lithium in Journal of Chemical Physics, 6, 367 (1938).

The essence of the $\mathbf{k} \cdot \mathbf{p}$ bandstructure model then is to identify a point in the $\mathbf{k}$-space where the electron eigenvalues and corresponding eigenfunctions are known, and to employ the machinery of perturbation theory to recover the values close to that $\mathbf{k}$-point. By including an appropriate number of starting states, this procedure is capable of recovering the eigenvalues and eigenfunctions in the *entire* Brillouin zone. For semiconductors, the $\mathbf{k}$-points of high importance are where the valence band maxima and conduction band minima occur. Since physical properties such as the energy bandgap (E_g), band-edge effective masses $(m_c^\star, m_v^\star)$, and spin-orbit splitting energy (Δ) that are intimately related to the eigenvalues and corresponding eigenfunctions at these band-edge $\mathbf{k}$-points are typically *known* from experiment, they constitute the "inputs" for $\mathbf{k} \cdot \mathbf{p}$ bandstructure calculations[2].

In Chapter 7 we discussed that there are two approaches to solve a perturbation problem in quantum mechanics. The most general approach is to construct the $\mathbf{k} \cdot \mathbf{p}$ matrix with a chosen restricted basis set, say containing the conduction band and valence band states at $\mathbf{k} = 0$. Use of symmetry considerations aids the diagonalization and solution of the matrices. In this chapter we begin the discussion of the $\mathbf{k} \cdot \mathbf{p}$ bandstructure method using the *analytical* perturbation theory to highlight several features and sum rules of bands in a simplified manner. Then we discuss the matrix approach, which is better suited to handle the degeneracies of the valence band of semiconductors, and to introduce spin-orbit coupling effects.

For an analytical, or "non-degenerate" $\mathbf{k} \cdot \mathbf{p}$ theory of semiconductor bandstructure, we recap the relevant results from Chapter 7. According to the **Brillouin–Wigner** (BW) theory, a perturbation $\hat{W}$ changes the eigenfunction from the unperturbed value $|u\rangle \rightarrow |\psi\rangle$ and the eigenvalue from the unperturbed value $E_u \rightarrow E$. The perturbed values are given in terms of the unperturbed ones by Equations 7.22 and 7.23 rewritten here to prepare for their application in $\mathbf{k} \cdot \mathbf{p}$ theory:

$$|\psi\rangle \approx |u\rangle + \underbrace{\sum_{m \neq u} \frac{\langle m|\hat{W}|u\rangle}{E - E'_m}|m\rangle}_{\phi^{(1)}}, \text{ and}$$

$$E \approx E_u + \underbrace{\langle u|\hat{W}|u\rangle}_{\Delta E^{(1)}} + \underbrace{\sum_{m \neq u} \frac{|\langle m|\hat{W}|u\rangle|^2}{E - E'_m}}_{\Delta E^{(2)}}. \tag{12.2}$$

Recall that the perturbed energy eigenvalue E is not completely solved by writing Equation 12.2 since E also appears in the denominator of the second order term $\Delta E^{(2)}$ on the right side. We will see later in this chapter that this feature can explain *non-parabolicity* of bands in narrow-bandgap semiconductors, and is also capable of handling degenerate energies, such as those of valence bands at $\mathbf{k} = 0$. Let us choose the unperturbed state at $\mathbf{k} = 0$ in the n^{th} band with energy $E_u = E_n(0)$. Inserting the $\mathbf{k} \cdot \mathbf{p}$ form of $\hat{W}$ from Equation 12.1, we

[2] Modern first-principles methods for the calculation of bandstructure can also provide the eigenvalues and eigenfunctions with sufficient accuracy. See Chapter 13 for further discussion.

obtain the central result of analytical $\mathbf{k}\cdot\mathbf{p}$ bandstructure theory:

$$\boxed{E_n(\mathbf{k}) \approx E_n(0) + \underbrace{\langle n|\frac{\hbar}{m_e}\mathbf{k}\cdot\mathbf{p} + \frac{\hbar^2k^2}{2m_e}|n\rangle}_{\Delta E^{(1)}} + \underbrace{\sum_{m\neq n}\frac{|\langle m|\frac{\hbar}{m_e}\mathbf{k}\cdot\mathbf{p} + \frac{\hbar^2k^2}{2m_e}|n\rangle|^2}{E_n(\mathbf{k}) - E_m(0)}}_{\Delta E^{(2)}}.} \tag{12.3}$$

This equation gives the bandstructure of band n by the $\mathbf{k}\cdot\mathbf{p}$ method[3]. The total second-order perturbation $\Delta E^{(2)}$ arises due to the "interaction" between different eigenvalues[4]. For matrix elements $\langle n|\hat{W}|m\rangle = 0$, there is no interaction. If $\langle n|\hat{W}|m\rangle \neq 0$, small energy energy separations $E_n(\mathbf{k}) - E_m(0)$ amplify the interaction. Whether $\langle n|\hat{W}|m\rangle$ vanishes or not is inferred from *symmetry* properties of the eigenfunctions and the perturbing potential W. This is the subject of the next section.

Before ending this section, we peek ahead to the result that if symmetry requires $\langle n|\mathbf{p}|n\rangle = 0$, since $\langle m|n\rangle = \delta_{mn}$, Equation 12.3 simplifies to

$$E_n(\mathbf{k}) - E_n(0) \approx \frac{\hbar^2k^2}{2m_e}[1 + \sum_{m\neq n}\frac{\frac{2|\mathbf{p}_{mn}|^2}{m_e}}{E_n(\mathbf{k}) - E_m(0)}], \tag{12.4}$$

where $\mathbf{p}_{mn} = \langle u_m(\mathbf{k}=0,\mathbf{r})|\hat{\mathbf{p}}|u_n(\mathbf{k}=0,\mathbf{r})$ is the momentum matrix element between the Bloch *lattice* functions. We shall find that the term in the numerator $E_P \equiv 2|\mathbf{p}_{mn}|^2/m_e \approx 20$ eV for several semiconductors of interest. If we consider just two states at $\mathbf{k} = 0$: the conduction band minimum $E_n(0) = E_c$ and a valence band maximum $E_m(0) = E_v$, with an energy bandgap of $E_c - E_v = E_g$, then since $E_g << 2|p_{cv}|^2/m_e$, the conditions for the **Rayleigh–Schrödinger** (RS) approximation apply. Replacing $E_c(\mathbf{k}) \to E_c$ in the second order correction, we get

$$E_c(\mathbf{k}) - E_c(0) \approx \frac{\hbar^2k^2}{2m_c^\star}, \text{ where } \frac{1}{m_c^\star} = \frac{1}{m_e}[1 + \frac{E_P}{E_g}]. \tag{12.5}$$

By connecting the conduction band-edge effective mass $m_c^\star$ to the bandgap, this result explains why small bandgap semiconductors exhibit small conduction band edge effective masses. An experimental measurement of the energy bandgap E_g and the conduction band-edge effective mass $m_c^\star$ yields the value of the momentum matrix element p_{cv} via this relation. For example, for GaAs, $E_g = 1.4$ eV and $m_c^\star = 0.067m_e$ yields $E_P = 2|p_{cv}|^2/m_e \approx 19.7$ eV and $p_{cv} \approx 1.2 \times 10^{-24}$ kg·m/s. Equating this to $\hbar k_0$ yields $k_0 \approx 2\pi/a_0 = G$, where $a_0 \approx 5.5$ Å, close to the lattice constant of GaAs (5.65 Å), and G is the reciprocal lattice constant. This is no coincidence; the conduction band edge states originate from the splitting of band degeneracy formed at $k = 0$ in the reduced zone scheme by the two nearly free–electron bands corresponding to $+G$ and $-G$, both of which possess crystal momentum of $|\hbar G|$ at $\mathbf{k} = 0$, as indicated in Fig. 12.1. We discuss aspects of symmetry in the next section, and use them to extend the discussion of the $\mathbf{k}\cdot\mathbf{p}$ bandstructure models to develop several similar insights of semiconductor bandstructure in the sections thereafter.

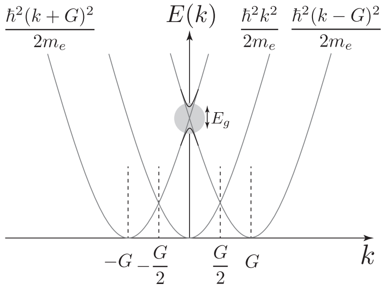

$\frac{\hbar^2(k+G)^2}{2m_e}$
$E(k)$
$\frac{\hbar^2k^2}{2m_e}$
$\frac{\hbar^2(k-G)^2}{2m_e}$
E_g
$-G$
$-\frac{G}{2}$
$\frac{G}{2}$
G
k

Fig. 12.1 Nearly free–electron origins of the band gap and band edges. In magnitude, the resulting momentum matrix elements have $p_{cv} \approx \hbar G$ and therefore $2p_{cv}^2/m_e \approx 2(\hbar^2G^2)/m_e \approx 20$ eV as a reasonable approximation for a lattice constant of $a_0 \approx 5.5$ Å. The relation $\langle c|\hat{p}|v\rangle \approx \hbar G$ means that the momentum operator "extracts" the *total* momentum of the state it acts on!

[3] $\mathbf{k}\cdot\mathbf{p} = -i\hbar(k_x\frac{\partial}{\partial x} + k_y\frac{\partial}{\partial y} + k_z\frac{\partial}{\partial z})$ in rectangular coordinates. The matrix elements therefore factorize into $\langle m|\mathbf{k}\cdot\mathbf{p}|n\rangle = \mathbf{k}\cdot\langle m|\mathbf{p}|n\rangle = \mathbf{k}\cdot\mathbf{p}_{mn}$ since $\mathbf{k}$ is not an operator.

[4] This "interaction" is, of course, of a purely mathematical origin. It results from splitting the full Hamiltonian $\hat{H}$ to an exactly solved part $\hat{H}^0$ and a perturbation $\hat{W}$. No new physical force is responsible for this "interaction"!

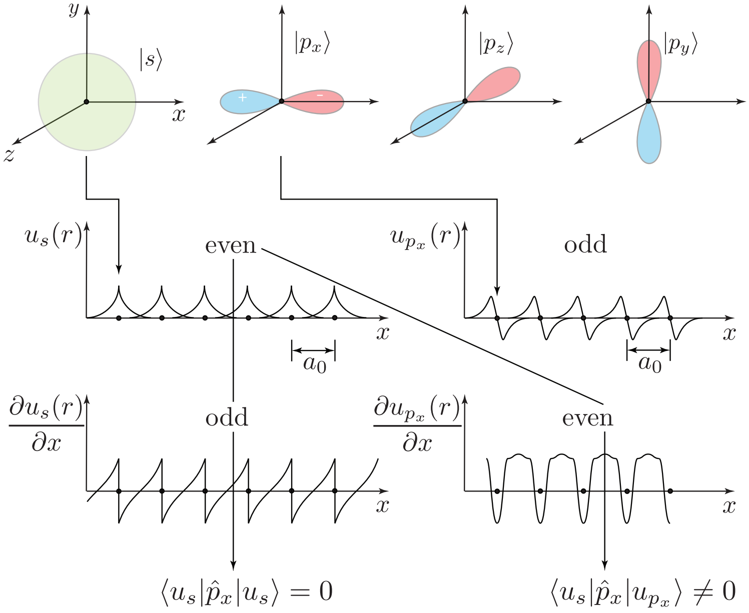

y
x
z
$|s\rangle$
$|p_x\rangle$
+
-
$|p_z\rangle$
$|p_y\rangle$
$u_s(r)$
even
$u_{p_x}(r)$
odd
x
a_0
x
a_0
$\frac{\partial u_s(r)}{\partial x}$
odd
$\frac{\partial u_{p_x}(r)}{\partial x}$
even
x
x
$\langle u_s|\hat{p}_x|u_s\rangle = 0$
$\langle u_s|\hat{p}_x|u_{p_x}\rangle \neq 0$

Fig. 12.2 s- and p-orbitals in atomic systems and in crystals. The s-orbital is spherical, and hence symmetric along all axes; the p-orbitals are antisymmetric or odd along the direction they are oriented – i.e., the p_x orbital has two lobes – one positive, and the other negative.

12.2 Symmetry

Symmetry considerations facilitate determining the momentum matrix elements of the kind $\mathbf{p}_{mn} = \langle u_m|\mathbf{p}|u_n\rangle$, where $\mathbf{p} = -i\hbar\nabla$ is the momentum operator, and u_n, u_m are the Bloch lattice functions at the $\mathbf{k}$ point chosen as the exactly solved (or unperturbed) problem. This $\mathbf{k}$-point typically is chosen as the $\mathbf{k} = 0$ or the Γ point for direct-bandgap semiconductors, but it does not have to be. The Bloch lattice function has the same periodicity as the lattice. We have seen in Chapter 11 in the tight-binding model that the Bloch lattice functions are the electron wavefunctions of the outermost electron orbitals participating in bonding, translated for every unit cell, shown schematically as $u_s(r)$ for the spherically symmetric s-orbitals and as $u_{p_x}(r)$ for the dumbbell shaped p_x-orbitals in Fig. 12.2.

In direct-gap semiconductors, for states near the conduction-band minimum ($\mathbf{k} = 0$), the Bloch lattice-function $u_c(\mathbf{k}, \mathbf{r}) = u_c(0, \mathbf{r})$ is predominantly $u_s(r)$, possessing even symmetry in every unit cell around a lattice site[5]. The states at the valence band maxima $u_v(r)$ are linear combinations of $u_{p_x}(r)$, $u_{p_y}(r)$, and $u_{p_z}(r)$. They have the odd symmetry around a lattice site and the axis of the p-orbitals, and rotational symmetry around its axis as seen in Fig. 12.2 for $u_{p_x}(r)$. Fig. 12.2 also schematically shows the derivatives of the Bloch lattice functions u_s and u_{p_x} in the x-direction. Note that the derivative turns a function with even symmetry around a lattice point into one with a odd symmetry, and vice versa. Since the momentum operator $\hat{p}_x = -i\hbar\frac{\partial}{\partial x}$ is odd in x, any momentum matrix element of the type $\langle \text{even}(x)|\frac{\partial}{\partial x}|\text{even}(x)\rangle$ and $\langle \text{odd}(x)|\frac{\partial}{\partial x}|\text{odd}(x)\rangle$ must be zero by symmetry[6].

[5] If the semiconductor has *indirect* bandgap, the conduction-band minimum state is no longer entirely $|s\rangle$-like; it has mixed $|s\rangle$ and $|p\rangle$ characteristics.

[6] Invoking periodicity, the matrix element is zero for the whole crystal if it is zero inside a unit cell.

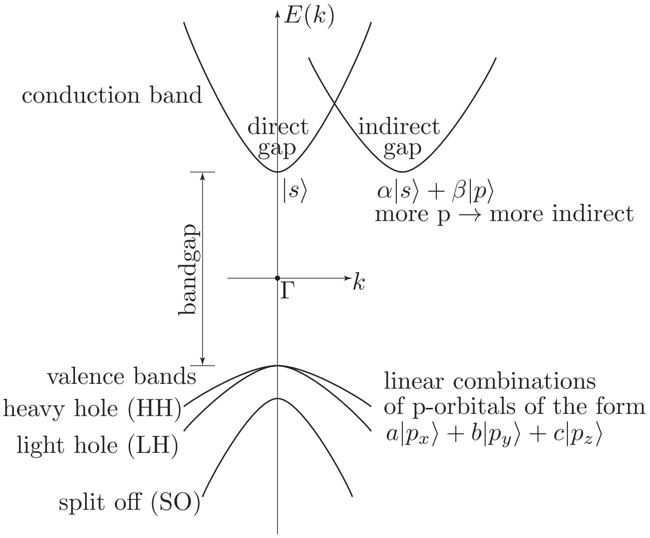


Fig. 12.3 The typical bandstructure of semiconductors. For direct-gap semiconductors, the conduction band state at $\mathbf{k} = 0$ is s-like. The valence band states are linear combinations of p-like orbitals. For indirect-gap semiconductors on the other hand, even the conduction band minima states have some amount of p-like nature mixed into the s-like state.

Fig. 12.3 summarizes the orbital origins of the band-edge states and the relative ordering of the bands around the bandgap of typical semiconductors. At $\mathbf{k} = 0$, the matrix element between the conduction band state Bloch lattice function and *any* valence band Bloch lattice function is

$$\langle u_c | u_v \rangle = 0, \tag{12.6}$$

i.e., it vanishes. This is seen from Fig. 12.2; the p states are odd along one axis and even along two others; however, the s states are even. So, the product, integrated over all unit cell is zero[7].

Consider next the momentum matrix element $\langle u_c|\mathbf{p}|u_v\rangle$ between the conduction and valence bands. Writing the valence band states as linear combinations of the type $|u_v\rangle = a|p_x\rangle + b|p_y\rangle + c|p_z\rangle$ where normalization requires $|a|^2 + |b|^2 + |c|^2 = 1$, we get

$$\langle u_c|\mathbf{p}|u_v\rangle = \langle s|\mathbf{p}(a|p_x\rangle + b|p_y\rangle + c|p_z\rangle)$$
$$= -i\hbar\langle s|(\hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z})(a|p_x\rangle + b|p_y\rangle + c|p_z\rangle)$$
$$= -i\hbar[\hat{x}a\langle s|\frac{\partial}{\partial x}|p_x\rangle + \hat{y}b\langle s|\frac{\partial}{\partial y}|p_y\rangle + \hat{z}c\langle s|\frac{\partial}{\partial z}|p_z\rangle]$$
$$\implies \boxed{\mathbf{p}_{cv} = \langle u_c|\mathbf{p}|u_v\rangle = [\hat{x}a + \hat{y}b + \hat{z}c]p_{cv},} \text{ and } |\mathbf{p}_{cv}| = p_{cv} \text{ where}$$
$$p_{cv} \equiv -i\hbar\langle s|\frac{\partial}{\partial x}|p_x\rangle = -i\hbar\int u_s(r)\frac{\partial u_{p_x}(r)}{\partial x}d^3r, \tag{12.7}$$

which is precisely the momentum matrix element mentioned in Equation 12.5 and estimated to be $p_{cv} \approx \hbar G$ where $G = 2\pi/a_0$, a reciprocal lattice vector[8]. In arriving at this result, we have made generous use of the fact that six of the nine terms that result from the matrix element are zero by symmetry. The reader is encouraged to verify this claim.

[7] The reader will notice that this is merely a visual confirmation: we already know that because these states have energies that are eigenvalues of Equation 12.1, the corresponding eigenfunctions are guaranteed to be orthogonal. The orthogonality holds not just for $\mathbf{k} = 0$, but for all $\mathbf{k}$.

[8] By symmetry we also have $p_{cv} = -i\hbar\langle s|\frac{\partial}{\partial y}|p_y\rangle = -i\hbar\langle s|\frac{\partial}{\partial z}|p_z\rangle$. In this chapter we assume that p_{cv} is known from experiment, as discussed following Equation 12.5.

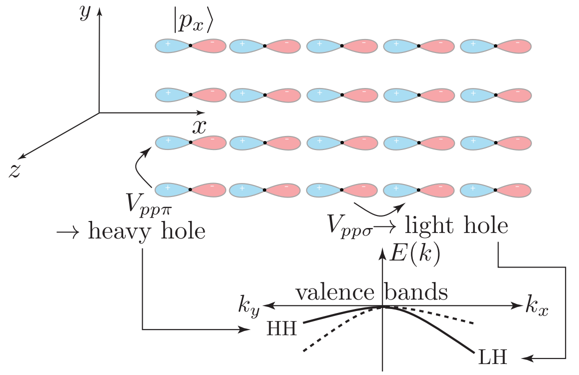


Fig. 12.4 Heavy and light hole bands near the top of the valence band formed from $|p_x\rangle$ orbitals attached to lattice sites. The light hole band along the k_x-axis is primarily due to the $V_{pp\sigma}$ bonding between the lobes of adjacent orbitals in the x-direction. The heavy-hole band along the k_y-axis is primarily due to the $V_{pp\pi}$ bonding between the lobes along the y-direction.

The general procedure to find the linear combination coefficients (a, b, c) for the valence bands is described in Section 12.5. It involves solving the corresponding $\mathbf{k} \cdot \mathbf{p}$ matrix eigenvalue problem, whereby the corresponding eigenfunctions yield the desired coefficients. Fig. 12.4 provides a simplified physical guide to the origin of the valence bands in a cubic lattice by neglecting spin. Following the discussion of tight-binding hopping and band curvatures in Chapter 11, we can argue that it is easier for electrons to hop between adjacent $|p_x\rangle$ orbitals along the x-direction (or with a wavevector in the k_x direction) than in the y-direction. Thus, the hopping parameters $V_{pp\sigma}$ are large and lead to a valence band that is more curved along the k_x axis due to the $|p_x\rangle$ orbitals. Since the effective masses go inversely as the hopping energy[9], this is the **light hole** (LH) band along k_x as shown in Fig. 12.4 as a solid line. Electrons can also hop between $|p_x\rangle$ orbitals in the y-direction, though it is not as easy as in the x-direction. This is modeled via a smaller hopping parameter $V_{pp\pi}$, leading to a band that is less curved, and has a larger effective mass. This valence band is the **heavy hole** (HH) band shown as a solid line in the k_y direction. In a 3D arrangement with $|p_x\rangle, |p_y\rangle$, and $|p_z\rangle$ orbitals at each lattice site, such $V_{pp\pi}$ hopping is also allowed between $|p_z\rangle$ orbitals in the y-direction. Therefore there are two HH bands along k_y. Thus, in the absence of spin-orbit coupling, three valence bands (one LH and two HH) form at each $\mathbf{k}$, and are degenerate at $\mathbf{k} = 0$. We can now apply the insights of this section to develop $\mathbf{k} \cdot \mathbf{p}$ theory.

12.3 Analytical model without spin

Using Fig. 12.5, we shall now extend the simple analytical bandstructure model of Equation 12.5 by including the new insights into the valence band structure. There is not one, but three valence bands that are degenerate at $\mathbf{k} = 0$. Therefore, the conduction band eigenvalues

---

[9] See Table 11.1 for the relations of bandwidths and effective masses to the hopping parameters.

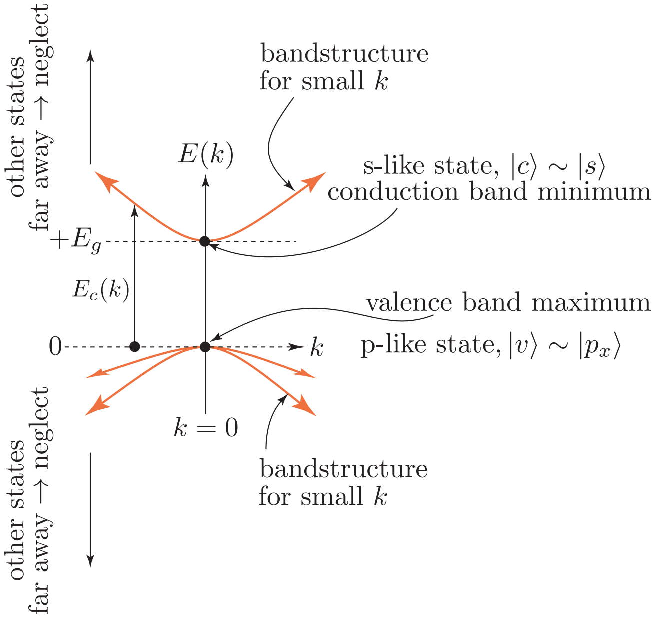

bandstructure
for small k
$E(k)$
s-like state, $|c\rangle \sim |s\rangle$
conduction band minimum
other states
far away → neglect
$+E_g$
$E_c(k)$
valence band maximum
0
k
p-like state, $|v\rangle \sim |p_x\rangle$
$k=0$
bandstructure
for small k
other states
far away → neglect

Fig. 12.5 The $\mathbf{k}\cdot\mathbf{p}$ bandstructure in the absence of spin-orbit coupling. The valence band edge is set to 0 of energy, and the conduction band edge is a band gap energy $+E_g$ above. The conduction bandstructure $E_c(k)$ is positive with a minimum of E_g, and the valence bandstructure $E_v(k)$ is negative with a maximum of zero.

should be given by

$$E_c(\mathbf{k}) \approx E_c(0) + \frac{\hbar^2 k^2}{2m_e} + \sum_{i=1}^{3} \frac{\frac{\hbar^2}{m_e^2}|\mathbf{k}\cdot\mathbf{p}_{ci}|^2}{E_c(\mathbf{k}) - E_i(0)}, \tag{12.8}$$

where $i = LH, HH1, HH2$ are the indices for the three valence bands, and $E_{HH1}(0) = E_{HH2}(0) = E_{LH}(0) = 0$ is chosen as the reference of energy. Each valence band state is a linear combination of the type $a_i|p_x\rangle + b_i|p_y\rangle + c_i|p_z\rangle$. To proceed further, we would need the individual coefficients of each of the three valence bands, for which we postpone discussion to Section 12.5. Considering the cubic symmetry, the interband matrix elements $\mathbf{p}_{c,LH}$, $\mathbf{p}_{c,HH1}$, and $\mathbf{p}_{c,HH2}$ are mutually orthogonal vectors, each of magnitude p_{cv}. Aligning them along cartesian axes, the wavevector $\mathbf{k}$ makes angles (α, β, γ) with the axes. Since $|\mathbf{k}\cdot\mathbf{p}_{ci}|^2 = k^2 p_{cv}^2 \cos^2\theta_i$, and the theorem for direction cosines $\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1$ holds. Writing $\varepsilon_0(k) = \frac{\hbar^2 k^2}{2m_e}$, Equation 12.8 becomes

$$E_c(k) \approx E_g + \frac{\hbar^2 k^2}{2m_e} + \frac{\frac{\hbar^2}{m_e^2} k^2 p_{cv}^2}{E_c(k)} = E_g + \varepsilon_0(k)[1 + \frac{E_P}{E_c(k)}]$$

$$\implies E_c(k) \approx [\frac{E_g + \varepsilon_0(k)}{2}] + \sqrt{[\frac{E_g + \varepsilon_0(k)}{2}]^2 + \varepsilon_0(k) E_P}, \tag{12.9}$$

where we have only retained the positive root of the quadratic equation for the conduction band. This form is clearly not parabolic in k for all k, a topic we take up in the next section. Only for small k and

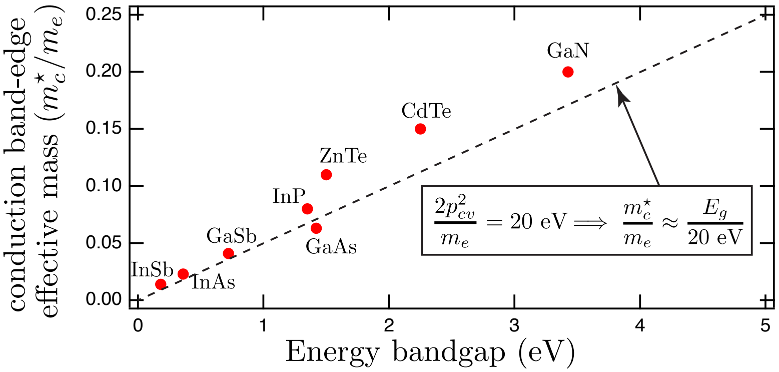


Fig. 12.6 Conduction band effective masses predicted from $\mathbf{k} \cdot \mathbf{p}$ theory. Note that the straight line is an approximate version of the result of $\mathbf{k} \cdot \mathbf{p}$ theory, and it does a rather good job for all semiconductors.

for $\varepsilon_0(k) << E_g$ we get

$$E_c(k) - E_g \approx \frac{\hbar^2 k^2}{2m_e}[1 + \frac{E_P}{E_g}\{1 - (\frac{\varepsilon_0(k)}{E_g}) + (\frac{\varepsilon_0(k)}{E_g})^2 - ...\}], \qquad (12.10)$$

from where by neglecting the terms of order $\varepsilon_0(k)/E_g$ and higher, we get the $k \to 0$, or conduction band edge effective mass as

$$E_c(k) \approx E_g + \frac{\hbar^2 k^2}{2m_c^\star}, \text{ where } \frac{m_e}{m_c^\star} = 1 + \frac{E_P}{E_g} \implies \boxed{m_c^\star = \frac{m_e}{1 + \frac{E_P}{E_g}} \approx \frac{E_g}{E_P} m_e.} \qquad (12.11)$$

This relation for the conduction band edge effective mass $m_c^\star$ is identical to Equation 12.5, but now obtained by accounting for the three valence bands. Fig. 12.6 shows in a dashed line a plot of this effective mass as a function of the bandgaps for various direct bandgap semiconductors. A good agreement is found for the trend if we choose $E_P = 2p_{cv}^2/m_e \approx 20$ eV. The reason for this value of E_P was also discussed following Equation 12.5.

Fig. 12.7 illustrates the effect of the bandgap on the curvature of the conduction band and therefore on $m_c^\star$. The closer are the energy eigenvalues for $\mathbf{k} = 0$ and for small $\mathbf{k}$ nearby, the smaller is the denominator $E_c(\mathbf{k}) - E_v(0)$ of the second order energy correction $\Delta E^{(2)}$ in the BW theory. Since this term is positive, $E_c(k)$ curves up sharply because of level repulsion from the valence band close by. This large curvature results in a small $m_c^\star$. Similarly, a larger bandgap semiconductor has weaker interband repulsion, less curvature of the band, and therefore a larger effective mass. Though this trend is borne out well in experiments in Fig. 12.6, it should be remembered that it is valid only for energies close to the bottom of the conduction band, as long as other bands (e.g. higher conduction bands other than the one CB and three VBs considered) are sufficiently far away in energy. Even if other bands are far in energy, for higher energy states of the conduction band, the parabolic approximation fails, which we now discuss.

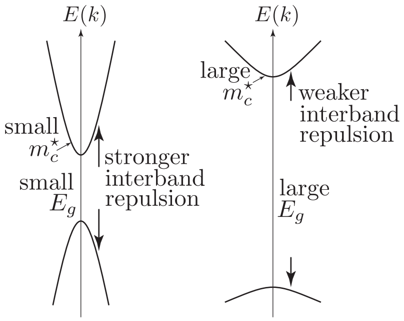


Fig. 12.7 Schematic drawing showing that a small energy bandgap results in strong interband repulsion and the large band curvature results in a small $m_c^\star$. Similarly for large energy bandgaps, weak interband repulsion results in a small curvature and hence a large $m_c^\star$.

12.4 Non-parbolicity and sum rules

From the general relation of the conduction band given by Equation 12.9, writing $\varepsilon_c(k) = E_c(k) - E_g$ as the excess conduction band energy above the band-edge, the reader can show that using the conditions $\varepsilon_0(k) << E_g$ and $m_c^\star << m_e$, the following approximation holds:

$$E_c(k) \approx [\frac{E_g + \varepsilon_0(k)}{2}] + \sqrt{[\frac{E_g + \varepsilon_0(k)}{2}]^2 + \varepsilon_0(k) E_P} \implies$$

$$\boxed{\varepsilon_c(k) \cdot [\varepsilon_c(k) + E_g] \approx \frac{\hbar^2 k^2}{2m_c^\star} \cdot E_g,} \text{ or } \boxed{\varepsilon_c(k) \cdot [1 + \gamma \varepsilon_c(k)] \approx \frac{\hbar^2 k^2}{2m_c^\star},} \tag{12.12}$$

where $\gamma = 1/E_g$ is referred to as a non parabolicity parameter. Fig. 12.8 schematically shows what is indicated in Equation 12.12 for the conduction band. For energies $\varepsilon_c(k) << E_g$ very close the bottom of the conduction band, the parabolic approximation characterized by the effective mass $m_c^\star$ is accurate.

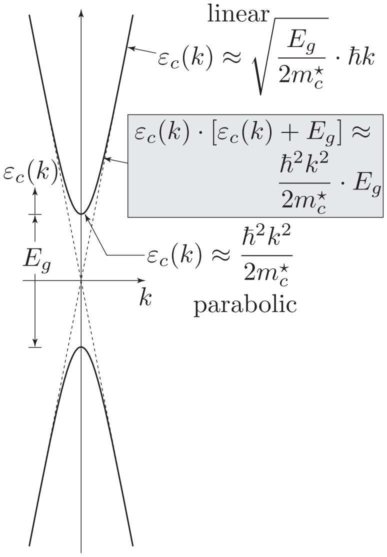

linear
$\varepsilon_c(k) \approx \sqrt{\frac{E_g}{2m_c^\star}} \cdot \hbar k$
$\varepsilon_c(k) \cdot [\varepsilon_c(k) + E_g] \approx \frac{\hbar^2 k^2}{2m_c^\star} \cdot E_g$
$\varepsilon_c(k)$
E_g
$\varepsilon_c(k) \approx \frac{\hbar^2 k^2}{2m_c^\star}$
k
parabolic

Fig. 12.8 Quantitative aspects of non–parabolicity of bands explained by the $\mathbf{k} \cdot \mathbf{p}$ model. Note how the band shape changes from parabolic near the conduction band minimum to linear far above it.

As $\varepsilon_c(k) \to E_g$, the band becomes increasingly non-parabolic, transitioning to a **linear** dispersion for $\varepsilon_c(k) >> E_g$. For such energies, it is seen from Equation 12.12 that $\varepsilon_c(k) \approx \hbar\sqrt{E_g/(2m_c^\star)} \cdot k$, indicating *linear* dependence on k. Furthermore, from Equation 12.11 since $m_c^\star \approx m_e \cdot (E_g/E_P)$, the linear dispersion becomes $\varepsilon_c(k) \approx \hbar v_b k$, where $v_b = \sqrt{E_P/(2m_e)} \approx 1.3 \times 10^8$ cm/s is the group velocity characterizing the linear energy dispersion. Note that this linear limit of group velocity is seemingly *independent* of the bandgap and effective mass, and depends on E_P which is similar for most semiconductors[10].

Consider a case of N bands when the energy levels at $\mathbf{k} = 0$ are well separated from each other with no degeneracies. In that case, applying the Rayleigh–Schrödinger approximation, we get that the effective mass at $\mathbf{k} = 0$ for band n, and its sum over all bands is

$$\frac{m_e}{m_n^\star} = 1 + \sum_{m \neq n} \frac{E_p}{E_n(0) - E_m(0)} \implies \boxed{\sum_n \frac{m_e}{m_n^\star} = N.} \tag{12.13}$$

This is proven by realizing that the energy difference denominators inside the sum occur in pairs of opposite signs when summed over all n, and therefore cancel. This is called the **effective mass sum rule**. This rule is a special case of a far more powerful sum rule called the **oscillator strength** or **F-sum rule**, which we simply state here without proof[11]. Instead of the Bloch lattice functions, the oscillator strength along the x-axis is defined in terms of the momentum matrix elements between the full Bloch function $|m, \mathbf{k}\rangle$, $|n, \mathbf{k}\rangle$, and their eigenvalues $E_m(\mathbf{k})$, $E_n(\mathbf{k})$ at any $\mathbf{k}$ in the Brillouin zone (not just at $\mathbf{k} = 0$):

$$F_x^{m,n} = \frac{\frac{2|\langle m,\mathbf{k}|\hat{p}_x|n,\mathbf{k}\rangle|^2}{m_e}}{E_m(\mathbf{k}) - E_n(\mathbf{k})} \implies \boxed{\sum_{m \neq n} F_x^{m,n} + \frac{m_e}{\hbar^2} \frac{\partial^2 E_n(\mathbf{k})}{\partial k_x^2} = 1.} \tag{12.14}$$

---

[10] This has a nice formal analogy to a relativistic, light-like particle moving with the group velocity v_b instead of the speed of light. For narrow bandgap semiconductors such as InSb, the Fermi level can be raised to this linear bandstructure region by field effect or heavy doping. Then the effects of non-parabolicity and the linear dispersion and high group velocity are experimentally observed (and exploited!).

[11] The F-sum rule applies to interband optical transitions in light-matter interaction problems. It is discussed in Chapter 27.

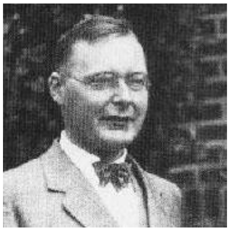


Fig. 12.9 Hans Kramers made significant contributions to quantum theory. He introduced the spin-degeneracy of electron states (Kramers' doublet). His name is associated with WKB tunneling theory (Chapter 24), and Kramers-Kronig relations in optical properties (Chapter 26).

12.5 The Kane model with spin-orbit interaction

The reader undoubtedly has noticed that we refrained from discussing the bandstructure of the valence bands in the last section. This is primarily due to their degeneracy. Though the Brillouin–Wigner form can be used to study them, in this section, we switch to the matrix form. In the process we also introduce spin-orbit interaction effects on the bandstructure. We will see that the introduction of spin-orbit interaction will rearrange the valence band energy levels, and split one of the three degenerate valence bands into a split-off (SO) band. The popular bandstructure model we shall discuss is due to Kane. He introduced it in 1956 to explain experimental observations of non parabolicity and optical transitions in semiconductors that motivated the need for properly including spin-orbit splitting in the valence band[12].

What is spin-orbit interaction? Since the topic is discussed in Chapter 25, Section 25.7 in more detail, a heuristic introduction sufficient for the Kane model is provided here. Overall, spin-orbit interaction is a *relativistic* effect, which implies the speed of light c will make an appearance! An electron in the outermost orbital with non zero angular momentum moving around the nucleus at relativistic speed $\mathbf{v}$ experiences an electric field $\mathbf{E} = (-\nabla V)/(-q)$, where V is the potential of the positively charged nucleus (+core electrons) as a Lorentz-transformed magnetic field $\mathbf{B}_{SO}$ given by

$$\mathbf{B}_{SO} = -\frac{1}{2}\frac{(\mathbf{v}\times\mathbf{E})/c^2}{\sqrt{1-\frac{v^2}{c^2}}} \approx -\frac{1}{2}\frac{\mathbf{v}\times\mathbf{E}}{c^2}, \tag{12.15}$$

where the approximation is for $v \ll c$. In the reference frame of the electron, the positively charged nucleus (+core electrons) rotates around it, forming a current loop which produces a magnetic field for the electron[13].

The effect of this internal magnetic field due to spin-orbit interaction is the introduction of an extra energy term in the Hamiltonian of the Schrödinger equation that is dependent on the electron spin, *even in the absence of external magnetic fields*. Each new state will still retain Kramers' double degeneracy (Fig. 12.9) in the absence of an external $\mathbf{B}$ field, but the allowed energies will be modified by the spin orbit interaction. For an electron of spin $\mathbf{s} = \mu_B\vec{\sigma}$ where $\mu_B = \frac{q\hbar}{2m_e}$ is the Bohr magneton and $\vec{\sigma} = \sigma_x\hat{x} + \sigma_y\hat{y} + \sigma_z\hat{z}$ is composed of the three Pauli spin matrices[14], the extra term in the Hamiltonian in a magnetic field $\mathbf{B}_{SO}$ is $\hat{H}_{SO} = \mathbf{B}_{SO}\cdot\mathbf{s}$, which becomes[15]

$$\hat{H}_{SO} = -\frac{\hat{\mathbf{v}}\times\mathbf{E}}{2c^2}\cdot\mathbf{s} = \frac{\hbar}{4m_e^2c^2}[\nabla V\times(\hat{\mathbf{p}}+\hbar\mathbf{k})]\cdot\vec{\sigma} \approx \frac{\hbar}{4m_e^2c^2}[\nabla V\times\hat{\mathbf{p}}]\cdot\vec{\sigma}. \tag{12.16}$$

Spin-orbit splitting effect is therefore observed in valence band electron states that originate from p-orbitals with non-zero orbital angular

[12] E. O. Kane, J. Phys. Chem. Solids, **1**, 82 (1956). Kane made pioneering contributions in the understanding of bandstructure, optical properties, and interband tunneling phenomena in semiconductors.

[13] How large is the spin-orbit magnetic field? Consider the model of a hydrogen atom. The velocity of electron in the ground state is $v \approx \alpha c$ where $\alpha = \frac{1}{137}$ is the fine structure constant. The consequent magnetic field seen by such an electron rotating at a Bohr radius $a_B^0 = 0.53$ Å from the nucleus is ... hold your breath ... $|\mathbf{B}_{SO}| \approx 6$ Tesla! That is a *very* large field, and should have perceivable effects.

[14] See Exercise 6.3, and later in this section.

[15] Here we have used the velocity operator $\hat{\mathbf{v}} = (\hat{\mathbf{p}} + \hbar\mathbf{k})/m_e$, and then used the fact that for $\mathbf{k} \approx 0$, $\hbar\mathbf{k} << \langle v|\hat{\mathbf{p}}|v\rangle \approx \hbar G$ for a momentum eigenstate, since the valence band states $|v\rangle$ have intrinsic "nearly free–electron" momentum $\hbar G$, where $G \sim 2\pi/a_0$ is the magnitude of the reciprocal lattice vector.

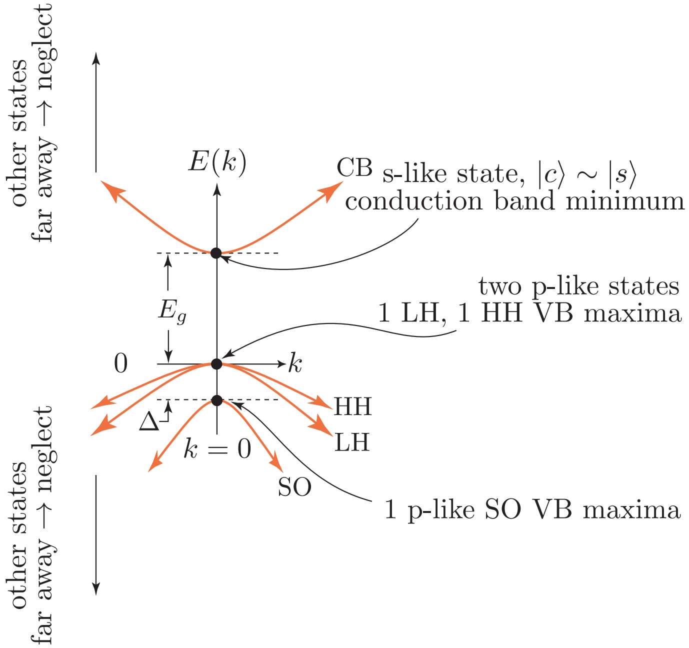


Fig. 12.10 $\mathbf{k} \cdot \mathbf{p}$ bandstructure with spin-orbit splitting Δ at $\mathbf{k} = 0$ in the valence band. The SO band is split off from the LH and HH bands due to spin-orbit interaction.

momentum. We will shortly introduce this term into the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian and find that the resulting bandstructure will take the form of Fig. 12.10. One of the three valence bands will be split off downward from the HH and LH bands, and form a separate (SO) band that is separated by an energy Δ at $\mathbf{k} = 0$. The Δ is called the **spin-orbit energy** and is experimentally measurable via optical transitions, and is shown in Fig. 12.11 for various semiconductors. Observe that the spin-orbit energy rapidly increases for semiconductors that have heavier atoms.

We can qualitatively argue that the magnitude of the spin-orbit splitting energy should be more for crystals whose constituent atoms have higher atomic number. This is because more protons in the nuclei produce higher electric fields. From perturbation theory we can predict that the spin-orbit splitting energy Δ of semiconductors should increases as the *fourth* power of the average atomic number of the constituent elements forming the semiconductor. This is indeed observed: Fig. 12.11 also shows the Z_{av}^4 polynomial fit, indicating a reasonable agreement.

Semiconductors made of lighter atoms from the top of the periodic table such as diamond, silicon, AlN and GaN have small Δ, and since they also have large energy bandgaps, $\Delta << E_g$ for them. The situation is switched for semiconductors such as InSb, InAs and GaSb made with heavier elements near the bottom of the periodic table. For the semiconductor InSb, $\Delta > E_g$, i.e., the spin-orbit splitting can substantially exceed the energy bandgap ($\Delta \sim 0.8$ eV, $E_g \sim 0.17$ eV at 300 K), by four times! We will now discuss Kane's bandstructure model that explained this experimentally observed fact.

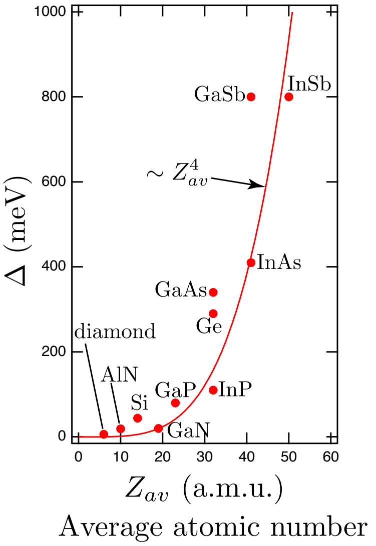


Fig. 12.11 The experimental spin-orbit splitting energy Δ_{so} for different semiconductors plotted against the average atomic number Z_{av}. It is a well-known result that the spin-orbit splitting for atomic systems goes as Z^4; the situation is not very different for semiconductors.

Adding the spin-orbit term from Equation 12.16 to the $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian of Equation 12.1 and rearranging, we get

$$\underbrace{(-\frac{\hbar^2}{2m_e}\nabla^2 + V}_{\hat{H}^0} + \underbrace{\frac{\hbar}{m_e}\mathbf{k}\cdot\hat{\mathbf{p}} + \frac{\hbar}{4m_e^2c^2}[\nabla V \times \hat{\mathbf{p}}]\cdot\vec{\sigma}}_{\hat{W}})u_{\mathbf{k}} = [E(\mathbf{k}) - \frac{\hbar^2k^2}{2m_e}]u_{\mathbf{k}}. \tag{12.17}$$

Note that we have taken the free–electron energy term $\frac{\hbar^2k^2}{2m_e}$ to the right side, and re-defined the perturbation term $\hat{W}$ to include the $\mathbf{k}\cdot\hat{\mathbf{p}}$ *and* the spin-orbit terms as the perturbation. At $\mathbf{k} = 0$, the unperturbed, or "solved" problem $\hat{H}^0 u_{\mathbf{k}=0} = E(\mathbf{k} = 0)u_{\mathbf{k}=0}$ has the following eight eigenvector and corresponding eigenvalues when the spin of each state is explicitly considered. The s orbital states of opposite spins $|S\uparrow\rangle$ and $|S\downarrow\rangle$ each have unperturbed energy E_s. The three up spin p orbital states $|X\uparrow\rangle, |Y\uparrow\rangle, |Z\uparrow\rangle$, and the corresponding three down spin states $|X\downarrow\rangle, |Y\downarrow\rangle, |Z\downarrow\rangle$ each have unperturbed energy[16] E_p. With this setup, we will now show that the Kane 8×8 $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian for the eight states is given by:

$$\begin{array}{cc}
 & \begin{array}{cccccccc} |S\uparrow\rangle & |S\downarrow\rangle & |X\uparrow\rangle & |Y\uparrow\rangle & |Z\uparrow\rangle & |X\downarrow\rangle & |Y\downarrow\rangle & |Z\downarrow\rangle \end{array} \\
\begin{array}{c} \langle S\uparrow| \\ \langle S\downarrow| \\ \langle X\uparrow| \\ \langle Y\uparrow| \\ \langle Z\uparrow| \\ \langle X\downarrow| \\ \langle Y\downarrow| \\ \langle Z\downarrow| \end{array} &
\begin{pmatrix}
E_s & 0 & k_xP & k_yP & k_zP & 0 & 0 & 0 \\
0 & E_s & 0 & 0 & 0 & k_xP & k_yP & k_zP \\
k_xP & 0 & E_p & -i\frac{\Delta}{3} & 0 & 0 & 0 & \frac{\Delta}{3} \\
k_yP & 0 & i\frac{\Delta}{3} & E_p & 0 & 0 & 0 & -i\frac{\Delta}{3} \\
k_zP & 0 & 0 & 0 & E_p & -\frac{\Delta}{3} & i\frac{\Delta}{3} & 0 \\
0 & k_xP & 0 & 0 & -\frac{\Delta}{3} & E_p & i\frac{\Delta}{3} & 0 \\
0 & k_yP & 0 & 0 & -i\frac{\Delta}{3} & -i\frac{\Delta}{3} & E_p & 0 \\
0 & k_zP & \frac{\Delta}{3} & i\frac{\Delta}{3} & 0 & 0 & 0 & E_p
\end{pmatrix}
\end{array}, \tag{12.18}$$

where in addition to the diagonal energy terms, the following are constant parameters in the Kane Hamiltonian:

$$\begin{aligned}
P &\equiv \frac{\hbar p_{cv}}{m_e} = \frac{\hbar}{m_e}\langle S|\hat{p}_x|X\rangle = \frac{\hbar}{m_e}\langle S|\hat{p}_y|Z\rangle = \frac{\hbar}{m_e}\langle S|\hat{p}_z|Z\rangle, \text{ and} \\
\Delta &\equiv \frac{3i\hbar}{4m_e^2c^2}\langle X|\frac{\partial V}{\partial x}\hat{p}_y - \frac{\partial V}{\partial y}\hat{p}_x|Y\rangle = \frac{3i\hbar}{4m_e^2c^2}\langle Y|\frac{\partial V}{\partial z}\hat{p}_y - \frac{\partial V}{\partial y}\hat{p}_z|Z\rangle \\
&= \frac{3i\hbar}{4m_e^2c^2}\langle Z|\frac{\partial V}{\partial z}\hat{p}_x - \frac{\partial V}{\partial x}\hat{p}_z|X\rangle.
\end{aligned} \tag{12.19}$$

The diagonal matrix elements of the Kane Hamltonian E_s and E_p are, by definition, solutions of the unperturbed Hamiltonian $\hat{H}^0$ and therefore for them $\langle ...|\hat{W}|...\rangle = 0$. Clearly, these would be the eigenvalues for $\mathbf{k} = 0$ *and simulataneously* when there is no spin-orbit coupling ($\Delta = 0$). This model was discussed in the last section, producing the sixfold degeneracy at $\mathbf{k} = 0$ for two spins of each state. We now evaluate a few off–diagonal matrix elements as examples, and encourage the reader to verify the rest. In this process, we will make use of the properties of Pauli spin matrices from Table 12.1, and the following expansion of the spin-orbit term in the Hamiltonian:

---

[16] Note the lowercase subscript for E_p to distinguish from the uppercase subscript of $E_P = 2p_{cv}^2/m_e$. Also note that we are using the notation $|p_x\rangle \equiv |X\rangle$ and so on.

$$\frac{\hbar}{4m_e^2c^2}[\nabla V \times \hat{\mathbf{p}}] \cdot \vec{\sigma} = \frac{\hbar}{4m_e^2c^2}[(\frac{\partial V}{\partial y}\hat{p}_z - \frac{\partial V}{\partial z}\hat{p}_y)\sigma_x + (\frac{\partial V}{\partial z}\hat{p}_x - \frac{\partial V}{\partial x}\hat{p}_z)\sigma_y + (\frac{\partial V}{\partial x}\hat{p}_y - \frac{\partial V}{\partial y}\hat{p}_z)\sigma_z]. \tag{12.20}$$

Table 12.1 Properties of spin matrices.

$\lvert\uparrow\rangle = \begin{bmatrix}1\\0\end{bmatrix}$	$\lvert\downarrow\rangle = \begin{bmatrix}0\\1\end{bmatrix}$
$\langle\uparrow\rvert = [1\ 0]$	$\langle\downarrow\rvert = [0\ 1]$
$\langle\uparrow\vert\uparrow\rangle = 1$	$\langle\downarrow\vert\downarrow\rangle = 1$
$\langle\uparrow\vert\downarrow\rangle = 0$	$\langle\downarrow\vert\uparrow\rangle = 0$
$\sigma_x = \begin{bmatrix}0 & 1\\1 & 0\end{bmatrix}$	$\sigma_x\lvert\uparrow\rangle = \lvert\downarrow\rangle$
	$\sigma_x\lvert\downarrow\rangle = \lvert\uparrow\rangle$
$\sigma_y = \begin{bmatrix}0 & i\\-i & 0\end{bmatrix}$	$\sigma_y\lvert\uparrow\rangle = i\lvert\downarrow\rangle$
	$\sigma_y\lvert\downarrow\rangle = -i\lvert\uparrow\rangle$
$\sigma_z = \begin{bmatrix}0 & 1\\-1 & 0\end{bmatrix}$	$\sigma_z\lvert\uparrow\rangle = \lvert\uparrow\rangle$
	$\sigma_z\lvert\downarrow\rangle = -\lvert\downarrow\rangle$

The matrix element $(1,3)$ is:

$$\begin{aligned}\langle S\uparrow\vert\hat{W}\vert X\uparrow\rangle &= \underbrace{\frac{\hbar}{m_e}\mathbf{k}\cdot\langle S\vert\hat{\mathbf{p}}\vert X\rangle}_{k_xP}\underbrace{\langle\uparrow\vert\uparrow\rangle}_{=1} + \frac{\hbar}{4m_e^2c^2}\langle S\uparrow\vert[\nabla V\times\hat{\mathbf{p}}]\cdot\vec{\sigma}\vert X\uparrow\rangle \\ &= k_xP + \frac{\hbar}{4m_e^2c^2}\times\Big(\langle S\uparrow\vert(\frac{\partial V}{\partial y}\hat{p}_z - \frac{\partial V}{\partial z}\hat{p}_y)\sigma_x\vert X\uparrow\rangle + \\ &\langle S\uparrow\vert(\frac{\partial V}{\partial z}\hat{p}_x - \frac{\partial V}{\partial x}\hat{p}_z)\sigma_y\vert X\uparrow\rangle + \langle S\uparrow\vert(\frac{\partial V}{\partial x}\hat{p}_y - \frac{\partial V}{\partial y}\hat{p}_x)\sigma_z\vert X\uparrow\rangle\Big) \\ &= k_xP + \frac{\hbar}{4m_e^2c^2}\times\Big(\langle S\vert\frac{\partial V}{\partial y}\hat{p}_z - \frac{\partial V}{\partial z}\hat{p}_y\vert X\rangle\underbrace{\langle\uparrow\vert\sigma_x\vert\uparrow\rangle}_{=0} + \\ &\langle S\vert\frac{\partial V}{\partial z}\hat{p}_x - \frac{\partial V}{\partial x}\hat{p}_z\vert X\rangle\underbrace{\langle\uparrow\vert\sigma_y\vert\uparrow\rangle}_{=0} + \underbrace{\langle S\vert\frac{\partial V}{\partial x}\hat{p}_y - \frac{\partial V}{\partial y}\hat{p}_x\vert X\rangle}_{=0}\underbrace{\langle\uparrow\vert\sigma_z\vert\uparrow\rangle}_{=1}\Big) \\ &\implies \langle S\uparrow\vert\hat{W}\vert X\uparrow\rangle = k_xP.\end{aligned} \tag{12.21}$$

The $\langle S\vert...\vert X\rangle = 0$ in the last term is confirmed using odd/even symmetry of the integrands. Note how the Pauli spin matrices act only on the spin operators and not on the orbital parts. Similarly, for the matrix element $(3,4)$ we get:

$$\begin{aligned}\langle X\uparrow\vert\hat{W}\vert Y\uparrow\rangle &= \frac{\hbar}{m_e}\mathbf{k}\cdot\underbrace{\langle X\vert\hat{\mathbf{p}}\vert Y\rangle}_{=0}\underbrace{\langle\uparrow\vert\uparrow\rangle}_{=1} + \frac{\hbar}{4m_e^2c^2}\langle X\uparrow\vert[\nabla V\times\hat{\mathbf{p}}]\cdot\vec{\sigma}\vert Y\uparrow\rangle \\ &= \frac{\hbar}{4m_e^2c^2}\times\Big(\langle X\vert\frac{\partial V}{\partial y}\hat{p}_z - \frac{\partial V}{\partial z}\hat{p}_y\vert Y\rangle\underbrace{\langle\uparrow\vert\sigma_x\vert\uparrow\rangle}_{=0} + \\ &\langle X\vert\frac{\partial V}{\partial z}\hat{p}_x - \frac{\partial V}{\partial x}\hat{p}_z\vert Y\rangle\underbrace{\langle\uparrow\vert\sigma_y\vert\uparrow\rangle}_{=0} + \underbrace{\langle X\vert\frac{\partial V}{\partial x}\hat{p}_y - \frac{\partial V}{\partial y}\hat{p}_x\vert Y\rangle}_{\neq 0}\underbrace{\langle\uparrow\vert\sigma_z\vert\uparrow\rangle}_{=1}\Big) \\ &\implies \langle X\uparrow\vert\hat{W}\vert Y\uparrow\rangle = \frac{\hbar}{4m_e^2c^2}\langle X\vert\frac{\partial V}{\partial x}\hat{p}_y - \frac{\partial V}{\partial y}\hat{p}_x\vert Y\rangle = \frac{\Delta}{3i} = -i\frac{\Delta}{3}.\end{aligned} \tag{12.22}$$

Here we have used the definition of the spin-orbit energy Δ from Equation 12.19. The term marked non-zero survives since it is a product of two odd functions in x and two odd functions in y, the net function being even. All elements of Equation 12.18 are obtained similarly.

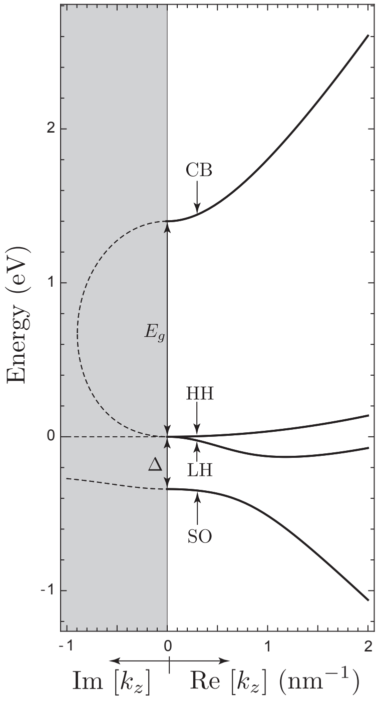


Fig. 12.12 Kane 8×8 bandstructure with spin-orbit interaction shown for real and imaginary k_z for the $(k_x, k_y) = (0, 0)$ direction. The parameters used are: $E_g = 1.4$ eV, $\Delta = 0.34$ eV, and P corresponding to $E_P = 2P^2/m_e = 20$ eV. As described in the text, $E_s = E_g$ and $E_p = -\Delta/3$ are used in the Kane matrix.

Before describing the analytical solutions of the Kane $\mathbf{k} \cdot \mathbf{p}$ matrix, let us investigate its numerical solution. Fig. 12.12 shows the calculated eigenvalues $E_n(\mathbf{k})$ of the Kane matrix (plus $\frac{\hbar^2 k^2}{2m_e}$ as required by Equation 12.17). The plot is for GaAs with input parameters E_g, P, and Δ described in the figure label.

Notice first the spin-orbit splitting Δ at $\mathbf{k} = 0$ of the valence band, and the formation of the split-off band. Each of the four bands are spin-degenerate. The real eigenvalues obtained for imaginary k_z are seen to connect the bands through the forbidden gap. An electron can follow this path when it tunnels from the valence to the conduction band. Interband tunneling is discussed in Chapter 24.

We will focus the remaining discussion in this chapter for bands for real k_z. While the bands look reasonable near $\mathbf{k} = 0$, the HH and LH bands are seen to curve *up* for large $\mathbf{k}$. This is because of the $\frac{\hbar^2 k^2}{2m_e}$ term, and indicates an unphysical situation brought about due to the approximations and neglect of other higher bands that push these eigenvalues down. These higher bands can be seen in the more accurate bandstructure of GaAs shown in Fig. 13.10, calculated by the empirical pseudopotential technique and confirmed by experiments. Therefore, the Kane model in the form presented here should be used for close to band-edge phenomena when carriers are limited to small $\mathbf{k}$. Inclusion of more bands can make the $\mathbf{k} \cdot \mathbf{p}$ bandstructure similar to that in Fig. 13.10 as shown in the texts referred to a the end of this chapter, and we do not pursue that approach. Instead, let us coax out the most information and insights from the simple Kane model before ending this chapter.

Note that putting $\mathbf{k} = 0$ in the Kane Hamiltonian makes the kP terms vanish, but the spin-orbit terms related to Δ remain. Thus, the matrix is not diagonal, and E_s and E_p are clearly not the energy eigenvalues at $\mathbf{k} = 0$. Finding the eigensystem (=eigenvalues+eigenfunctions) of the Kane matrix at $\mathbf{k} = 0$ reveals that the eigenvalues are given by two states at E_s, four states at $E_p + \frac{\Delta}{3}$, and two states at $E_p - \frac{2\Delta}{3}$. From Fig. 12.12 we identify the two E_s states to be the conduction band minima of opposite spins, the four $E_p + \frac{\Delta}{3}$ states as the maxima of the LH and HH bands double degeneracy each, and the two $E_p - \frac{2\Delta}{3}$ states as the SO band maximum of double degeneracy. Thus, the energy bandgap is $E_g = E_s - (E_p + \frac{\Delta}{3})$, and the spin-orbit splitting is $(E_p + \frac{\Delta}{3}) - (E_p - \frac{2\Delta}{3}) = \Delta$. Setting the LH and HH energies at $k = 0$ as the zero of energy, we get $E_p + \frac{\Delta}{3} = 0$, which leads to $E_s = E_g$, parameters that used for Fig. 12.12.

The eigenfunctions at $\mathbf{k} = 0$ indicate the mixing of the states due to spin-orbit coupling. The conduction band states corresponding to energy $E_s = E_g$ remain uncoupled from the valence band states and each other, as $|S \uparrow\rangle$ and $|S \downarrow\rangle$, as may be seen from the Kane matrix (Equation 12.18) by putting $k_x = k_y = k_z = 0$. Because of the mixing, the rest of the eigenstates are *not* $|X \uparrow\rangle$, $|Y \uparrow\rangle$ and so on as assumed for the Kane Hamiltonian in Equation 12.18, but may be found by re-

Table 12.2 Kane band eigenstates and bandstructure near $k = 0$.

$\|\psi\rangle$	$\|\psi(k=0)\rangle$	Notation $\|j, m_j\rangle$	$E(k=0)$	$E(k)$	Effective mass
$\|CB\uparrow\rangle$	$\|iS\uparrow\rangle$	-	E_g	$E_g + \frac{\hbar^2 k^2}{2m_c^\star}$	$\frac{m_e}{m_c^\star} = 1 + \frac{\frac{2}{3}E_P}{E_g} + \frac{\frac{1}{3}E_P}{E_g+\Delta}$
$\|CB\downarrow\rangle$	$\|iS\downarrow\rangle$	-	E_g	″	″
$\|HH\uparrow\rangle$	$-\frac{1}{\sqrt{2}}\|(X+iY)\uparrow\rangle$	$\|\frac{3}{2}, \frac{3}{2}\rangle$	0	$\frac{\hbar^2 k^2}{2m_e}$	m_e (inaccurate)
$\|HH\downarrow\rangle$	$\frac{1}{\sqrt{2}}\|(X-iY)\downarrow\rangle$	$\|\frac{3}{2}, -\frac{3}{2}\rangle$	0	″	″
$\|LH1\rangle$	$\frac{1}{\sqrt{6}}\|(X-iY)\uparrow\rangle + \sqrt{\frac{2}{3}}\|Z\downarrow\rangle$	$\|\frac{3}{2}, -\frac{1}{2}\rangle$	0	$\frac{\hbar^2 k^2}{2m_{LH}^\star}$	$\frac{m_e}{m_{LH}^\star} = \frac{\frac{2}{3}E_P}{E_g} - 1$
$\|LH2\rangle$	$-\frac{1}{\sqrt{6}}\|(X+iY)\downarrow\rangle + \sqrt{\frac{2}{3}}\|Z\uparrow\rangle$	$\|\frac{3}{2}, \frac{1}{2}\rangle$	0	″	″
$\|SO1\rangle$	$\frac{1}{\sqrt{3}}\|(X-iY)\uparrow\rangle - \frac{1}{\sqrt{3}}\|Z\downarrow\rangle$	$\|\frac{1}{2}, -\frac{1}{2}\rangle$	$-\Delta$	$-\Delta - \frac{\hbar^2 k^2}{2m_{SO}^\star}$	$\frac{m_e}{m_{SO}^\star} = \frac{\frac{2}{3}E_P}{E_g+\Delta} - 1$
$\|SO2\rangle$	$\frac{1}{\sqrt{3}}\|(X+iY)\downarrow\rangle + \frac{1}{\sqrt{3}}\|Z\uparrow\rangle$	$\|\frac{1}{2}, \frac{1}{2}\rangle$	$-\Delta$	″	″

diagonalizing the matrix for each desired $\mathbf{k}$. In particular, for the $\mathbf{k} = 0$ point, the linear combinations of the original states that form the new eigenfunctions after including *both* the $\mathbf{k} \cdot \mathbf{p}$ term and the spin-orbit terms in the Hamiltonian are shown in Table 12.2, along with their corresponding eigenvalues at $E(k = 0)$, and their extension into the individual bands for $k \neq 0$, which gives the respective effective masses of each band. The $|j, m_j\rangle$ refers to the angular momentum notation highlighting the mixing of the spin and orbital angular momentum due to spin-orbit interaction[17].

Several points can be now made about the effect of spin-orbit interaction. Note that the conduction band states as well as the heavy hole states are spin eigenstates of σ_z, but the light hole and split off states are not. The individual bandstructures are obtained from the secular equation of the 8×8 matrix. The effective masses of the resulting bands can therefore be obtained from the secular equation, as mentioned in Exercise 12.4. The effective mass of the heavy hole band is a ”sore point” in this model, as was already alluded to in Fig. 12.12, where it was seen to curve the wrong way even for small k! Rather than dwelling on this point, let us discuss the bandstructures of the LH and SO bands that have emerged from the Kane model, and their impact, via the spin-orbit interaction, on the conduction band.

Since momentum matrix elements of the type $\langle Y|\mathbf{p}|X\rangle$ and its cyclic permutations vanish by symmetry, it is clear that the SO band only couples to the CB. Using Rayleigh–Schrödinger theory, the bandstructure $-\Delta - \frac{\hbar^2 k^2}{2m_{SO}^\star}$ is obtained, with the factor $E_g + \Delta$ which is the energy separation between the coupled states at $\mathbf{k} = 0$. The effective mass for the LH band has a similar experssion, but corresponding to an energy separation of E_g.

We end this chapter by illustrating the effect of spin-orbit interaction on the conduction band edge effective mass. Since we know the

[17] The reader is encouraged to verify these results following the earlier discussion in the $|S\rangle$, $|X\rangle$, ... basis (see Exercise 12.4). In particular, it can be verified that the eigenfunctions are all orthogonal to each other, and are individually normalized.

eigenfunctions at $\mathbf{k} = 0$, we re-evaluate the conduction bandstructure by direct application of Rayleigh–Schrödinger theory:

$$
\begin{aligned}
E_{c\uparrow}(k) \approx E_g + \frac{\hbar^2 k^2}{2m_e} &+ \\
\frac{\hbar^2}{m_e^2}\Big[\frac{|\mathbf{k}\cdot\langle HH\uparrow|\hat{\mathbf{p}}|CB\uparrow\rangle|^2}{E_g} + \frac{|\mathbf{k}\cdot\langle HH\downarrow|\hat{\mathbf{p}}|CB\uparrow\rangle|^2}{E_g} &+ \\
\frac{|\mathbf{k}\cdot\langle LH1|\hat{\mathbf{p}}|CB\uparrow\rangle|^2}{E_g} + \frac{|\mathbf{k}\cdot\langle LH2|\hat{\mathbf{p}}|CB\uparrow\rangle|^2}{E_g} &+ \\
\frac{|\mathbf{k}\cdot\langle SO1|\hat{\mathbf{p}}|CB\uparrow\rangle|^2}{E_g+\Delta} + \frac{|\mathbf{k}\cdot\langle SO2|\hat{\mathbf{p}}|CB\uparrow\rangle|^2}{E_g+\Delta}\Big] & \\
\approx E_g + \frac{\hbar^2 k^2}{2m_e} + \frac{\hbar^2 p_{cv}^2}{m_e^2}\Big[\frac{\frac{1}{2}(k_x^2+k_y^2)}{E_g} + \frac{\frac{1}{6}(k_x^2+k_y^2)}{E_g} + \frac{\frac{2}{3}k_z^2}{E_g} &+ \\
\frac{\frac{1}{3}(k_x^2+k_y^2)}{E_g+\Delta} + \frac{\frac{1}{3}(k_z^2)}{E_g+\Delta}\Big] & \\
\implies E_{c\uparrow}(k) \approx E_g + \frac{\hbar^2 k^2}{2m_c^\star} \text{ where } \frac{m_e}{m_c^\star} = 1 + \frac{\frac{2}{3}E_P}{E_g} + \frac{\frac{1}{3}E_P}{E_g+\Delta}. &
\end{aligned}
\tag{12.23}
$$

Note that taking $\Delta \to 0$ in this expression gives back $\frac{m_e}{m_c^\star} = 1 + \frac{E_P}{E_g}$, which was discussed and illustrated in Fig. 12.6, one of the triumphs in insight afforded by the $\mathbf{k}\cdot\mathbf{p}$ theory!

12.6 Chapter summary section

In this chapter, we learned:

- Analytical formulation of the bandstructure is possible near the band edges with the energy eigenvalues at $\mathbf{k} = 0$ as the inputs via the $\mathbf{k}\cdot\mathbf{p}$ theory, and
- The $\mathbf{k}\cdot\mathbf{p}$ method of bandstructure calculation provides insight into the relationships between bandgaps and effective masses by band repulsion, potential transitions between bands via symmetries, spin-orbit interaction terms, and also provides an efficient way to treat heterostructures and lower-dimensional semiconductor quantum wells, wires, and dots.

Further reading

As Kittel states in his text on solid state physics, learning how to calculate bandstructure is an *art*, not learnt from book only, but by experience. **Fundamentals of Semiconductors** by Yu and Cardona: Chapter 2 in this comprehensive text has a modern treatment of semiconductor bandstructure and specially of the $\mathbf{k} \cdot \mathbf{p}$ theory. It makes heavy usage of group theory, which can be intimidating for beginners, but nevertheless very rewarding. The authors do not assume knowledge of group theory – they actually have "crystallized" the results that are needed from group theory in the chapter. **Energy Bands in Semiconductors** by Donald Long is an old and classic monograph. The theory is covered in 80 pages, and the rest of the book analyzes bandstructures of specific materials. Being out of print, it is difficult to get! A classic on the subject of bandstructure of semiconductors is **Band Structure of Semiconductors** by I. M. Tsidilkovski, with an excellent, comparative discussion of various techniques of bandstructure calculation. **Basic Semiconductor Physics** by C. Hamaguchi has a detailed description of the $\mathbf{k} \cdot \mathbf{p}$ bandstructure calculation method with examples of full-band evaluations for several semiconductors.

Exercises

(12.1) **Non parabolicity effects of semiconductor bandstructure**

Consider a narrow bandgap 3D semiconductor with a bandgap of $E_g = 0.17$ eV and conduction band-edge effective mass of $m_c^\star = 0.014 m_e$.

(a) An electron density of this heavily doped semiconductor is measured to be $n_{3d} = 10^{19}$ /cm^3. Assuming parabolic bands, calculate the location of the Fermi level with respect to the conduction band edge, i.e., compute $(E_F - E_c)$ in meV scale at $T = 300$ K.

The non parabolicity at high energies changes the density of states of the conduction band.

(b) Calculate the modified density of states (DOS) of the conduction band as a function of energy using the non-parabolic approximation in the final form of Equation 12.12:

$$\varepsilon_c(k)[\varepsilon_C(k) + E_g] = \frac{\hbar^2 k^2}{2 m_c^\star} E_g. \tag{12.24}$$

(c) Re-calculate the location of the Fermi level with the non-parabolic DOS. Compare with part (a). What is the error?

(d) Calculate and plot the group velocities as a function of energy for the parabolic and non-parabolic bands and sketch them. Argue why the asymptotic limit for the non-parabolic bands do not depend sensitively on the choice of semiconductor.

(e) Now re-derive (b) and (d) for the 2D and 1D cases. Can you give examples of materials in 2D and 1D where these considerations are of paramount importance?

(12.2) **Effect of spin-orbit interaction on effective masses**

Table 12.2 lists the effective masses of the CB, LH, and SO bands as a function of the spin-orbit energy Δ and the bandgap E_g, in addition to the standard interband momentum matrix element energy parameter E_P. Re-make the plot of Fig. 12.6 taking into account the spin-orbit splitting, and discuss the errors incurred if it is neglected. Make similar plots of the LH and SO effective masses of semiconductors, and discuss the accuracy of the 8×8 Kane $\mathbf{k} \cdot \mathbf{p}$ model for the valence band edge features.

(12.3) **Sum rules**

(a) Verify the "sum rule" of Equation 12.13 for effective masses at the Γ point $(k = 0)$. This implies that if I know effective masses of $N - 1$ bands, I can find that of the one unknown band. Apply it to Table 12.2 and discuss its applicability. Note that the effective mass in the sum rule uses the following sign convention instead of what we have used in this textbook: for downward bending bands, the effective mass is negative if the energy is written as $+\frac{\hbar^2 k^2}{2m^\star}$.

(b) Verify the F-sum rule of Equation 12.14.

(12.4) **Verifying properties of the Kane Hamiltonian**

(a) Verify the claim that the eigenfunctions and eigenvalues of the 8×8 Kane Hamiltonian are as shown in Table 12.2.

(b) Verify the bandstructures and the effective masses of each band shown in the table, by solv-

ing the secular equation.
(c) Verify the bandstructures and the effective masses of each band shown in the table, by applying the Rayleigh–Schrödinger perturbation theory. Treat the LH and HH bands by resorting to the Brillouin–Wigner perturbation theory.

(12.5) **Eight is not enough: extending $\mathbf{k} \cdot \mathbf{p}$**
The eight band Kane model fails to capture the details of the heavy hole band. Read in the references listed in the Further Reading section to find how extending the matrix by including higher and lower bands, the accuracy of the $\mathbf{k} \cdot \mathbf{p}$ model can be increased significantly, and bands far from $\mathbf{k} = 0$ can also be evaluated accurately. Of course this requires heavy use of more parameters and symmetry considerations. Nevertheless, larger sized matrices are routinely used for $\mathbf{k} \cdot \mathbf{p}$ bandstructure calculations in the design and analysis of bulk and quantized semiconductor heterostructures. One of the reasons is that it is simple to introduce effects such as strain in this model.

(12.6) **Crystal field splitting in wurtzite crystals**
In cubic crystals, we saw that spin-orbit interaction split off one valence band, but at $k = 0$, the LH and HH remained degenerate. This is no more the case in a wurtzite crystal which is uniaxial. Discuss why the LH and HH degeneracy is split at $k = 0$ for semiconductors such as GaN and discuss the values of the crystal-field splitting and the spin-orbit splitting in such crystals. In fact, a rather interesting reversal of the ordering of valence bands occurs for the ultra wide bandgap wurtzite semiconductor AlN. Read about this inversion, the reasons for it, the magnitude of the energy separations, and the impact this reversal of ordering causes in the optical properties of this semiconductor.

1, 2, 3 ... ∞: Pseudopotentials and Exact Bandstructure

13

13.1 The empire strikes back 269
13.2 Exact bandstructure of the Dirac comb potential 270
13.3 Tight-binding models emerge from Kronig–Penney 274
13.4 Point defects in Kronig–Penney models 275
13.5 Green's functions from Kronig–Penney models 277
13.6 Pseudopotentials: what they are and why they work 278
13.7 Bandstructure of Si, Ge, and GaAs 281
13.8 Bandstructure of AlN, GaN, and InN 283
13.9 Pseudopotentials to DFT and beyond 285
13.10 Chapter summary section 286
Further reading 287
Exercises 287

This chapter is intended to connect with the state of the art in bandstructure calculations of semiconductors. It brings you right up to the point where you can calculate full bands over the entire Brillouin zone and use it in research, using a laptop computer as your sole computational resource. At the same time, it connects to the more computationally intensive approaches. The goals are to understand and learn:

- The exact solutions of the bandstructure of crystalline solids.
- Pseudopotentials: what they are, why they work, and why they are so powerful.
- Example calculations of exact bandstructures of a few semiconductors (e.g. silicon, GaAs, GaN, MoS_2) on a laptop.
- Pseudopotentials as the precursor to density functional theory techniques for electronic bandstructure and beyond.

13.1 The empire strikes back

We studied the tight-binding method to calculate the electronic bandstructure of crystals in Chapter 11, and the $\mathbf{k} \cdot \mathbf{p}$ method in Chapter 12. Both methods provided deep insights into the properties of semiconductors such as orbitals and their hybridization, and the connection between effective masses and bandgaps. The forte of tight-binding models is an unmatched intuitive guide and an accurate representation of the valence bands. The $\mathbf{k} \cdot \mathbf{p}$ method provides most details of the bandstructure near the band edges E_c and E_v. However, some physical phenomena in semiconductors involve energy eigenvalues $E(k)$ of k-states that sample a large part, or **the entirety of the Brillouin zone**. Such processes may be the optical reflectivity over a large range of wavelengths, and high field transport phenomena such as impact ionization, high-field current saturation, and Gunn oscillations. To explain these physical phenomena, accurate energy eigenvalues of both the valence and the conduction bands over the entire BZ is desired, calling for an exact method: the subject of this chapter.

In this final chapter on electron bandstructure, we encounter one exactly solvable model of electron bandstructure based on the technique outlined in Chapter 9, Equation 9.8 that is rooted in the introduction of Bloch wavefunctions into the Schrödinger equation. The exactly solvable model is not of a real crystal, but of a hypothetical crystal, all of whose atoms introduce a Dirac delta function potential well for electrons. This exactly solved bandstructure model will be our passport to the **empirical pseudopotential** technique, which provides some of the most accurate electron bandstructures of semiconductors to date. We will then see how the concept of pseudopotentials leads naturally to the modern method of first-principles (or *ab initio*) calculations of the electronic bandstructure of crystals, that go under the generic name of density functional theory (DFT).

13.2 Exact bandstructure of the Dirac comb potential

Fig. 13.1 Ralph Kronig, with Penney in 1931, offered the first exact solution of the quantum mechanical states of an electron in a 1D periodic potential. With Hans Kramers (Fig. 12.9), he developed the relation between the real and imaginary parts of the dielectric constant, which go by the name of Kramers–Kronig relations.

In this section we discuss a non–perturbative or *exactly* solvable model of electron bandstructure in a crystal: the celebrated **Kronig–Penney model** (Fig. 13.1). The purpose of the solution is to illustrate much of bandstructure physics and also to develop a bag of useful concepts that permeate much of solid state physics. The Kronig–Penney model is an **exactly solvable periodic potential problem** in quantum mechanics for the electron. The problem is exactly solvable in all dimensions – we consider the 1D case. The periodic potential is modeled as a series of Dirac delta functions

$$V(x) = \sum_n S\delta(x - na), \tag{13.1}$$

where a is the lattice constant, and S is the strength of the potential. The sum over n runs over all lattice sites. For example, for a 1D closed ring of length L with $N = L/a$ lattice points and lattice constant a, $0 \leq n \leq N-1$. This is schematically represented in Fig. 13.2.

Using reciprocal lattice vectors $G_n = \frac{2\pi}{a}n$ with the identity $\sum_n \delta(x - na) = \sum_n \frac{1}{a} e^{-i\frac{2\pi}{a}nx}$ and substituting in the Schrödinger equation, we get[1]

$$[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{S}{a}\sum_n e^{-iG_n x}]\psi = E\psi. \tag{13.2}$$

The allowed wavefunctions are Bloch functions, which are Fourier expanded in $G_m = \frac{2\pi}{a}m$ as

$$\psi_k(x) = e^{ikx}u_k(x) = e^{ikx}\sum_m u_{G_m} e^{iG_m x} = \sum_m u_{G_m} e^{i(k+G_m)x}. \tag{13.3}$$

Note that $\psi(x=0) = \sum_{G_m} u_{G_m}$, the sum of *all* the Bloch coefficients, in other words, as long as G's are reciprocal lattice vectors, $\sum_G u_G =$

---

[1] We use a mass m here for generality – for electrons the free–electron mass should be used.

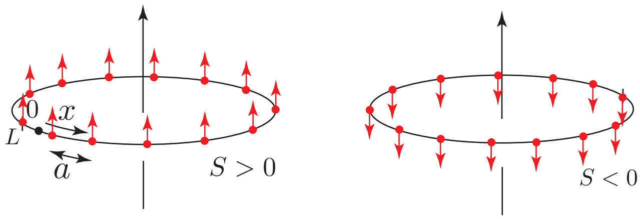


Fig. 13.2 The Kronig–Penney "Dirac" comb periodic potential for a particle on a ring. Left: positive or repulsive potential for $S > 0$, and Right: Negative or attractive potentials for $S < 0$.

$\psi(0)$. Substituting Equation 13.3 in the Schrödinger equation gives

$$\sum_m \frac{\hbar^2(k+G_m)^2}{2m} u_{G_m} e^{iG_m x} + \frac{S}{a}\sum_m\sum_n u_{G_m} e^{i(G_m-G_n)x} = E_k \sum_m u_{G_m} e^{iG_m x}. \tag{13.4}$$

Multiplying by e^{-iGx} and integrating over all x, we use to identity $\int_0^L e^{i(G'-G)x}dx = L\delta_{G,G'}$ to get

$$\frac{\hbar^2(k+G)^2}{2m} u_G + \frac{S}{a}\sum_n u_{G_n+G} = E_k u_G. \tag{13.5}$$

Solving for u_G yields

$$u_G = \frac{S}{a}\frac{\sum_{G_n+G} u_G}{E_k - \frac{\hbar^2(k+G)^2}{2m}}. \tag{13.6}$$

Now for a very useful trick: summing both sides over $G's$ cancels the u_G terms because $\sum_{G_n+G} u_G = \sum_G u_G = \psi(x=0)$, leaving us with the identity

$$\boxed{1 = \frac{S}{a}\sum_G \frac{1}{E_k - \frac{\hbar^2(k+G)^2}{2m}}.} \tag{13.7}$$

This is a rather fancy way of writing unity! Note that this is an *exact* form of the solution of Schrödinger's equation for the periodic potential problem. Inverting it into the form

$$\frac{a}{S} = \sum_G \frac{1}{E_k - \frac{\hbar^2(k+G)^2}{2m}}, \tag{13.8}$$

we are in a position to investigate the aftermath of the solution in Equation 13.7.

Fig. 13.3 shows a graphical solution of Equation 13.8 plotted as a function of the energy E_k for two values of k. When the strength of the potential $S > 0$, the LHS of Equation 13.8 is $a/S > 0$, and is the constant shown as the horizontal solid line in the figure. The RHS is a function of energy E_k, with a number of poles located at $E_k = \frac{\hbar^2(k+G)^2}{2m}$ (where the RHS diverges). There are several points of intersection –

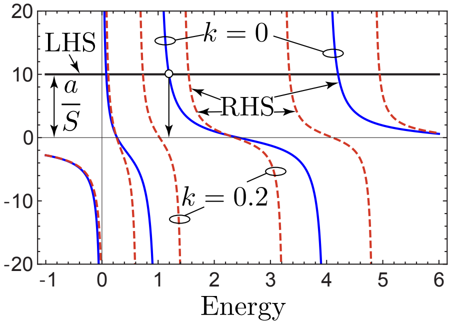


Fig. 13.3 A graphical solution scheme for the repulsive Kronig–Penney Dirac comb. Note that the lowest energy is larger than zero. LHS stands for the left hand side and RHS to the right hand side of Equation 13.8.

one of which is highlighted as a circle. The energies E_k corresponding to these intersection points are the only allowed eigenvalues for the problem. There are several allowed eigenvalues: in fact, there are exactly N distinct eigenvalues corresponding to $n = 0, 1, ..., N-1$ values of $G_n = \frac{2\pi}{a} n$ and that many poles.

If we turn the strength of the potential down by taking $S \to 0$, the solid horizontal line corresponding to a/S goes off to $+\infty$, and the intersections of the RHS and LHS then are exactly at the N energies for which the RHS blows up. Clearly, these energy eigenvalues are at $E_k = \frac{\hbar^2 (k+G)^2}{2m}$, and we have recovered the nearly free-electron model of the electron.

If the strength is made negative by letting $S < 0$, it is clear that the red line $a/S < 0$, and there is an energy intersection for energy that is *negative*, i.e., $E_k(min) < 0$.

Fig. 13.4 shows the calculated energy bandstructures $E(k)$ for $S > 0$ (left) and $S < 0$ (right). The axes are in units of $F = \frac{\hbar^2}{2m} \cdot (\frac{\pi}{a})^2$ for energy, and $\frac{2\pi}{a}$ for k. The solid lines in the figure represents several important features of *any* bandstructure in the presence of a non-zero periodic potential. This is superposed on the dashed line plot of bandstructure when the periodic potential is turned off ($S = 0$), but the electron wavefunction is still required to satisfy the lattice periodicity and symmetry, the "nearly" free–electron (NFE) model with $E = \frac{\hbar^2 (k+G)^2}{2m}$. Note that for a repulsive potential with $S > 0$, the Kronig–Penney bandstructure energies are *higher* than the NFE values at *all* values of k except at the Brillouin zone center and edges $k = 0, \pm\frac{\pi}{a} n$. The highest eigenvalue of each Kronig–Penney band is degenerate with the NFE eigenvalues of $E(k = n\frac{\pi}{a}) = n^2 \cdot F$, where $n = 1, 2, ...$, locating energy eigenvalues $F, 9F, ...$ at $k = \pm\frac{\pi}{a}$ at the BZ edge, and $4F, 16F, ...$ at $k = 0$ as the *maxima* of the corresponding bands.

That the energy eigenvalues for $S > 0$ are higher (or equal to) than the NFE values is *guaranteed* by the Hellmann–Feynman theorem. The Hellmann–Feynman theorem[2] states that the eigenvalues E_k of any Hamiltonian $\hat{H}$ satisfy $\frac{\partial E_k}{\partial \lambda} = \langle k | \frac{\partial \hat{H}}{\partial \lambda} | k \rangle$. Imagine the Kronig–Penney potential as a perturbation to the NFE Hamiltonian $\hat{H} = \hat{H}_0 + \lambda \hat{W}$ where $W(x) = S \sum_n \delta(x - na)$, and $\hat{H}_0 |k\rangle = E_k^0 |n\rangle$, the eigenvalues of the NFE

[2] Recall Section 7.5.

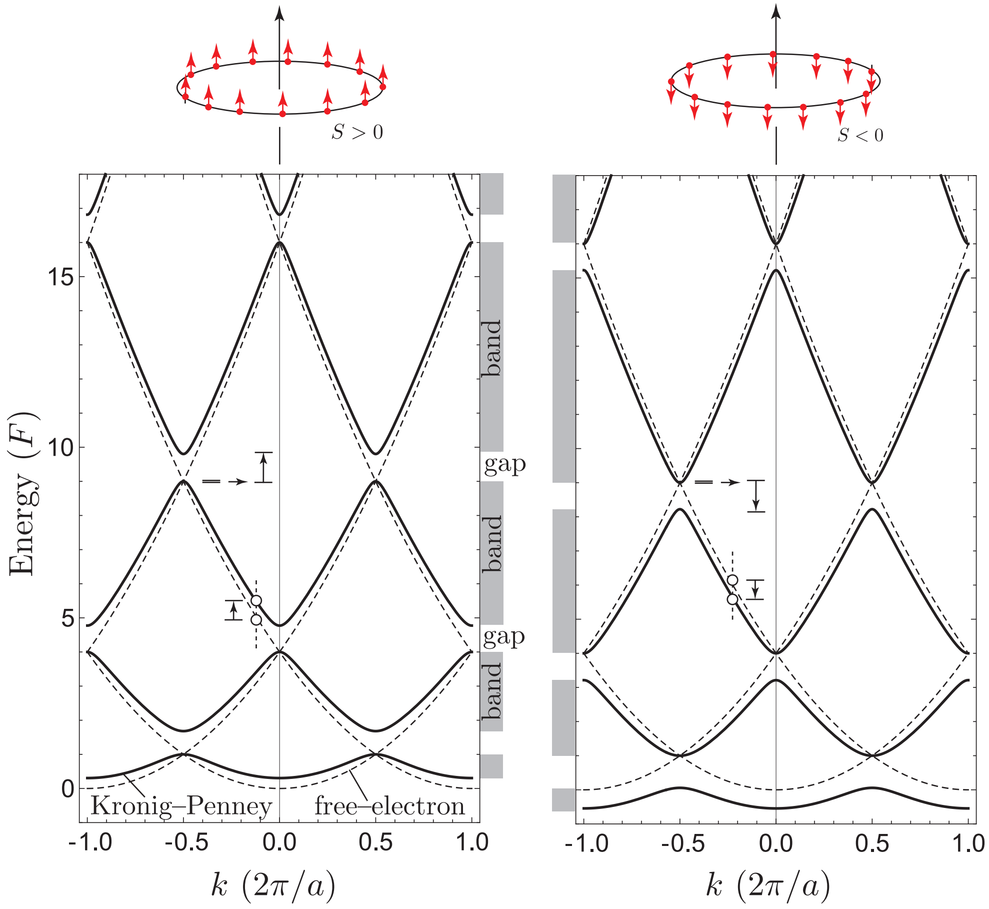


Fig. 13.4 The solid lines show the bandstructure for repulsive (left) and attractive (right) Kronig–Penney potentials. The nearly free–electron bandstructure $E(k) = \frac{\hbar^2(k+G)^2}{2m}$ is shown as dashed lines. The allowed energy bands are indicated in gray along with the energy gaps.

model $E_k^0 = \frac{\hbar^2(k+G)^2}{2m}$ shown by the dashed lines in Fig. 13.4. Then, we must have $\frac{\partial E_k}{\partial \lambda} = \langle k|\frac{\partial(\hat{H}_0+\lambda W)}{\partial \lambda}|k\rangle = \langle k|W|k\rangle = \int dx|\psi_k(x)|^2 W(x) = SN|u_k(0)|^2 \geq 0$, and the perturbed eigenvalue $E_k \geq E_k^0$. This remains true at all points in k-space except at points of degeneracy, as indicated by an arrow in the left figure of Fig. 13.4. At k-points were eigenvalues are degenerate, the splitting is such that for $S > 0$, one eigenvalue increases, while the other stays put. The lowest energy allowed is $E_{min}^+ > 0$ for $S > 0$, and the lowest band is rather narrow. This means the electron is ”sluggish” in this band, and it has a large effective mass. As we move up to higher energies, the points of degeneracy develop sharper curvatures and the bands become wider, making the electron effective mass lighter.

Note the differences for the attractive delta potential ($S < 0$) band structures highlighted by the right panel in Fig. 13.4, and drawn at exactly the same scale for easy comparison. The lowest energy allowed now is $E^{-}_{min} < 0$ for $S < 0$, i.e.. it is *negative* in stark contrast to the situation for $S > 0$. The Hellmann–Feynman theorem now guarantees that the eigenvalues are *lower* than the NFE case. At the k-points of degeneracy, the splitting is such that one eigenvalue stays put again, but the other is pushed *down*, exactly opposite to the case of $S > 0$. The lowest eigenvalue of each Kronig–Penney band is degenerate with the NFE eigenvalues of $E(k = n\frac{\pi}{a}) = n^2 \cdot F$ again, where $n = 1, 2, ...$, locating energy eigenvalues $F, 9F, ...$ at $k = \pm\frac{\pi}{a}$ at the BZ edge, and $4F, 16F, ...$ at $k = 0$ as now the *minima* of the corresponding bands.

13.3 Tight-binding models emerge from Kronig–Penney

In Chapter 11 we discussed the tight-binding model for bandstructure starting from atomic orbitals. We will now see that tight-binding models emerge naturally from the exact Kronig Penney model. Applying the trigonometric identity $\cot(x) = \sum_{-\infty}^{+\infty} \frac{1}{n\pi + x}$ to the right hand side of the central Kronig–Penney eigenvalue Equation 13.8, with $G_n = n\frac{2\pi}{a}$, Equation 13.8 transforms into:

$$\cos(ka) = \cos(qa) + \frac{mSa}{\hbar^2} \cdot \frac{\sin(qa)}{qa}, \tag{13.9}$$

where $q = \sqrt{2mE_k/\hbar^2}$. This is still an *exact* solution of the Schrödinger equation. *Now the values of E_k that satisfy this equation will form the energy bandstructure $E(k)$ for each k.* The left hand side is limited to $-1 \leq \cos(ka) \leq +1$, but the RHS of Equation 13.9 can reach values up to $1 + \frac{mSa}{\hbar^2} = 1 + C$ which can clearly exceed unity[3]. This restricts the allowed values of q for real energy eigenvalues $E = \frac{\hbar^2 q^2}{2m}$ for each k. Fig. 13.5 shows the "bands" of q where the RHS lies between $-1 \leq RHS \leq +1$, and real energy eigenvalues are allowed.

Now the zeroes of $\frac{\sin(x)}{x}$ occur at $x = n\pi$ where $n = \pm 1, \pm 2, ...$ It is clear that a band of q-values, and corresponding energies are allowed near the zeroes of the RHS as indicated in Fig. 13.5 (left). Let us find an approximate solution for the first band $E_1(k)$ by expanding the RHS for a large strength, or $C = \frac{mSa}{\hbar^2} >> 1$ near the first zero at $n = 1$, around $qa = \pi$. Using $\delta = \pi - qa$, the expansion yields $\cos(qa) + C \cdot \frac{\sin(qa)}{qa} \approx -1 + \frac{C}{\pi}\delta$, which when used in Equation 13.9 yields

$$E_1(k) \approx E_0 - 2J(1 + \cos ka), \tag{13.10}$$

where $E_0 = \frac{\pi^2\hbar^2}{2ma^2}$ coincides with the NFE energy at $k = \frac{\pi}{a}$, and the "hopping" or tunneling term is $J = \frac{\pi^2\hbar^4}{2m^2a^3S} = \frac{E_0}{C}$. This is clearly in the form of a tight-binding model of Chapter 11! Now we really don't

[3] The maximum value of $\frac{\sin x}{x}$ is 1.

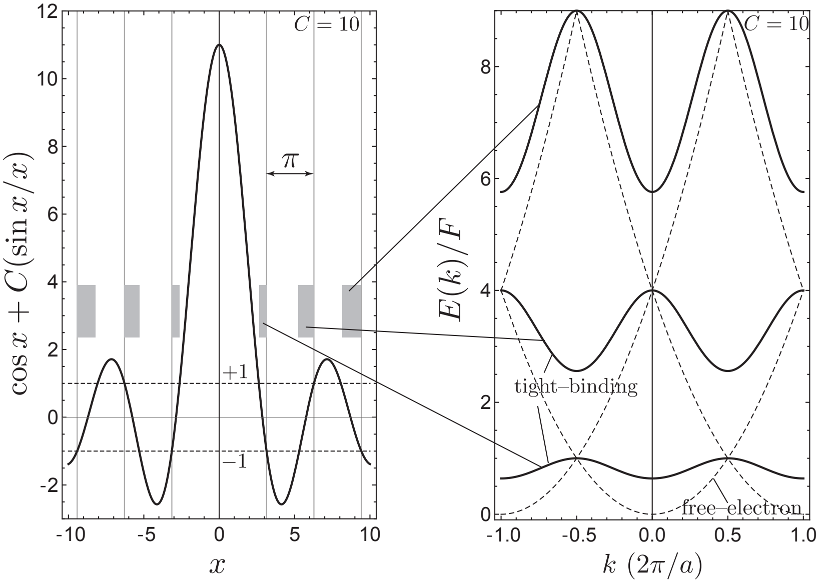


Fig. 13.5 The left figure shows a plot of the RHS of Equation 13.9 with $x = qa$, and the LHS is limited to $-1 \leq LHS \leq +1$. The narrow bands within which the two sides are equal are highlighted; each leads to an allowed energy band. Because the intersections are near $x = n\pi$ where $n = \pm 1, \pm 2, ...$, an approximate analytical expression of all the bands can be obtained (see Equation 13.11). This first three tight-binding bandstructures are plotted in the right panel. Compare with Fig. 13.4.

need to stop at the first root – expanding around $qa = n\pi$, and retaining only the linear expansion terms, we get a more comprehensive tight-binding bandstructure of the n^{th} band as:

$$E_n(k) \approx n^2 E_0 \big[1 - \frac{1}{C} + \frac{(-1)^n}{C}\cos(ka)\big]^2. \tag{13.11}$$

Fig. 13.5 shows a plot of the first three bands for the dimensionless strength $C = 10$. Note that the energy eigenvalues at the BZ edges co-incide with the free-electron values. This is similar to the case for $S > 0$ in the Kronig Penney model in Fig. 13.4 (Left).

13.4 Point defects in Kronig–Penney models

Now imagine that in the Kronig–Penney model, only *one* of the N sites has a potential that is different from the other sites. Let us call this difference in the strength U_0, meaning at this particular site, the delta-function strength is $S + U_0$ instead of S, where U_0 can be positive or negative. What is the effect on the energy eigenvalues and the eigenstates due to the presence of this "defect"?

This problem can now be solved because the exact solution of the Kronig–Penney model *without the defect* has given us the eigenvalues for each k-state in the BZ as $E_{KP}(k)$ – for example – shown in Fig. 13.4. Then, we go through exactly the same procedure that led to the Kronig–Penney solution in Equation 13.8, and end up with the new solution

$$\frac{Na}{U_0} = \sum_k \frac{1}{E_k - E_{KP}(k)}, \tag{13.12}$$

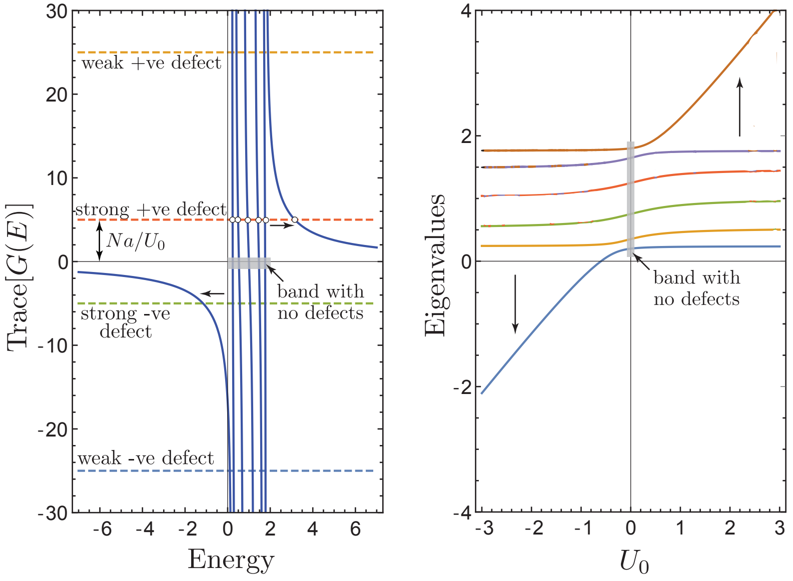


Fig. 13.6 Figures showing the effect of defect states on the allowed energy eigenvalues as a function of the defect potential strength. The left figure shows the graphical solution to the Kronig–Penney type solution, and in particular illustrates the splitting off of one eigenvalue – the highest eigenvalue of the band for positive defect potentials, and the lowest energy eigenvalues for negative defect potentials. This is further highlighted in the figure on the right, where the eigenvalue spectrum is plotted as a function of the defect potential.

where k are the allowed states in the first BZ, N is the number of lattice sites, and therefore $Na = L$ is the macroscopic length. Clearly, in the absence of the defect, $U_0 \to 0$, and the LHS$\to \infty$. This happens exactly N times in the RHS when the allowed energies $E_k = E_{KP}(k)$, i.e., we recover the original Kronig–Penney solution without the defect, as we should.

But when $U_0 \neq 0$, the allowed energies E_k must deviate from $E_{KP}(k)$ to satisfy Equation 13.12. To illustrate the solution graphically, we plot the RHS and the LHS in Fig. 13.6. We will see in the next section that the RHS of Equations 13.12 and 13.8 are actually the Trace of the Green's function matrix of the problem, i.e., $\sum_k \frac{1}{E_k - E_{KP}(k)} = \text{Trace}[\hat{G}(E)]$. The plot in Fig. 13.6 for a few-site chain shows the effect of the defect on the eigenvalue spectrum clearly. The figure on the right illustrates the movement of the eigenvalues as the strength of the defect U_0 is tuned from zero to large positive and large negative values. The eigenvalues at $U_0 = 0$ constitute the band without the defect. When the defect strength is +ve and strong, the LHS L/U_0 line moves closer to the x-axis (left figure), and it is clear that one of the intersections – at the *top* of the energy band splits off from the band rapidly, whereas all other eigenvalues do not change as much. Any change is positive, as guaranteed by the Hellmann–Feynman theorem. This is a characteristic feature – similarly, for a negative U_0, the lowest eigenvalue of the band splits off and leaves other eigenvalues mostly unchanged.

We will see later that $U_0 > 0$ "defects" explain the formation of acceptor states at the top of valence bands, and are designed such that the splitting energy is less than $k_b T$ for room-temperature generation of holes. Similarly, the bottom of the band with $U_0 < 0$ models donor states and electron doping at the bottom of the conduction band of semiconductors. This exact model also explains the formation of deep level defects inside the bandgap of semiconductors.

13.5 Green's functions from Kronig–Penney models

We noted the repeated appearance of sums over the Brillouin zone of the kind $\sum_k \frac{1}{E-E(k)}$ which have units of $(\text{energy})^{-1}$. This may be thought of as a function of the variable E, or energy. The reason why such sums permeate exact solutions of problems will now become clear: and will lead us to define Green's functions.

Consider the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi \rightarrow [i\hbar\frac{\partial}{\partial t} - \hat{H}]\Psi = 0. \tag{13.13}$$

Let us think of the equation as the product of the operator (or matrix) $i\hbar\frac{\partial}{\partial t} - \hat{H}$ with Ψ. For this product to be zero, either $i\hbar\frac{\partial}{\partial t} - \hat{H}$, or Ψ, or both must be zero. The only interesting case here is when we actually have a quantum object with a nonzero wavefunction, $\Psi \neq 0$. Thus, $i\hbar\frac{\partial}{\partial t} - \hat{H}$ should be zero. Now we have learned that if the quantum object is in a state of definite energy, $i\hbar\frac{\partial}{\partial t}\Psi_n = E_n\Psi_n$, Ψ_n, and E_n is a real eigenvalue representing the energy of the state. Let us generalize this and write $i\hbar\frac{\partial}{\partial t} = E$, where E is a *variable*. We can then write the Schrödinger equation as $[EI - \hat{H}]\Psi = 0$, where I is an identity operator, or the identity matrix when the equation is written out for any chosen basis. However, the equation in this form does not hold true for all E, but only for certain $E = E_n$ – only when the variable E matches up with an allowed eigenvalue. Now let us think of $EI - \hat{H}$ as a *function* of E. When we vary E, this function has very sharp responses when $E = E_n$: the function is a "detector" of eigenvalues – it detects an eigenvalue by vanishing. At those sharp energies, $\Psi = \Psi_n \neq 0$ is an eigenfunction, so the function provides the eigenfunction as its "residue". Now with this qualitative picture in mind, let us solidify the concept of the Green's function of the system.

We like detectors to "scream" when they detect, rather than to go silent. So, can we find a function $\hat{G}$ that instead of solving the equation $[EI - \hat{H}]\Psi = 0$, solves the equation $[EI - \hat{H}]\hat{G} = I$ instead? Formally, the function is clearly $\hat{G} = [EI - \hat{H}]^{-1}$. This function clearly blows up when $E = E_n$, and is indeed the screaming detector we are looking for. It is the Green's function for the Hamiltonian $\hat{H}$. Let us assume that we know all the eigenvalues of a particular Hamiltonian $\hat{H}_0$ to be E_n and the corresponding eigenfunctions are $|n\rangle$. The Green's function can then be written out as a matrix form:

$$\hat{G}_0(E) = \sum_n [EI - \hat{H}]^{-1}|n\rangle\langle n| = \sum_n \frac{|n\rangle\langle n|}{E - E_n}. \tag{13.14}$$

It is clear that the Green's function is actually a matrix, and sums of the kind that appeared earlier in the solution of the Kronig–Penney and the defect problems are the sum of the diagonal terms in a diagonal basis. Now it turns out that the sum of the diagonal terms is *invariant* to what basis is chosen for the matrix – which is why it goes by a name – the Trace. Thus, we have a very important relation

$$\text{Trace}[\hat{G}(E)] = \sum_k \frac{1}{E - E_0(k)}, \qquad (13.15)$$

where $E_0(k)$ are the allowed eigenvalues of the system. The solution of the Kronig–Penney model is thus very compactly written in the formal way as $\text{Trace}[\hat{G}_0(E)] = \frac{a}{S}$, where $\hat{G}_0(E) = (EI - \hat{H}_0)^{-1}$, and $\hat{H}_0|k\rangle = E_0(k)|k\rangle$ are the nearly free–electron eigensystem, with $E_0(k) = \frac{\hbar^2(k+G)^2}{2m}$. The solution of a single-site defect state of strength S_0 is then written as $\text{Trace}[\hat{G}(E)] = \frac{Na}{S_0}$, where now the Green's function is for the exactly solved Kronig–Penney eigensystem $\text{Trace}[\hat{G}(E)] = (EI - \hat{H})^{-1}$, where $\hat{H}|k\rangle = E_{KP}(k)|k\rangle$, and $E_{KP}(k)$ are the eigenvalues of the Kronig–Penney model.

We can write the Green's function in a non-diagonal basis as well. For the general basis states $|n\rangle$ and $|l\rangle$ we can write

$$\hat{G}_0(l,l';E) = \sum_n \langle l[EI - \hat{H}]^{-1}|n\rangle\langle n|l'\rangle = \sum_n \frac{\langle l|n\rangle\langle n|l'\rangle}{E - E_n}, \qquad (13.16)$$

which includes both diagonal and off-diagonal terms of the matrix. Since this form does not require us to start with a diagonal basis, it is preferred and is called the **spectral representation** of the Green's function. For example, this can be applied to the nearly free–electron "on a ring" of circumference L and lattice constant a. Since the eigenfunctions are $\langle x|n\rangle = \frac{1}{\sqrt{L}}e^{ik_n x}$ of energy $E_n = \frac{\hbar^2 k_n^2}{2m_e}$, its Green's function is

$$\hat{G}_0(x,x';E) = \frac{1}{L}\sum_n \frac{e^{ik_n(x-x')}}{E - E_n} = \frac{1}{L}\int_{-\frac{\pi}{a}}^{+\frac{\pi}{a}} \frac{dk}{\frac{2\pi}{L}} \frac{e^{ik_n(x-x')}}{E - \frac{\hbar^2 k^2}{2m_e}}, \qquad (13.17)$$

which depends on both the spatial dimension and energy. This form of the Green's function is related to the **local** density of states.

Fig. 13.7 Conyers Herring developed the orthogonalized plane-wave (OPW) method in 1940s, quite a bit ahead of the times. He created the theoretical division of Bell Laboratories, and led and influenced several solid–state theorists and experimentalists. He contributed to the understanding of electron transport in semiconductors: the ionized impurity scattering theory is described in Chapter 23. He identified accidental band crossings, which have today developed into topological Weyl semimetals. Herring also introduced several semiconductor jargon, such as "valleys", and "intervalley processes".

13.6 Pseudopotentials: what they are and why they work

The 3D version of the Dirac delta potential is $V(\mathbf{r}) = S\delta(\mathbf{r})$ where S has units of eV·cm^3. For any crystal lattice (e.g. FCC, HCP, etc), a 3D Dirac comb potential is formed by attaching a Dirac delta at each lattice point: $V(\mathbf{r}) = S\sum\delta(\mathbf{r} - \mathbf{R_n})$, where $\mathbf{R_n} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$ is a real space lattice vector. Assuming S to be negative, and a one atom basis, we can imagine it to be the attractive potential well introduced by the charges on the nucleus, and all the core electrons of that atom. For example, in a silicon crystal, there are four valence electrons, so the (nucleus + core electrons) have a net charge of $Q = +4e$. If the entire charge is assumed to be at the center of the nucleus (this is clearly a massive approximation!), the *unscreened* Coulomb potential will be $V(r) = -(4e^2)/4\pi\epsilon_0 r$, which is a deep potential well. The Dirac delta potential with a negative S is a very crude model of this potential. The Fourier transform of the 3D Dirac comb potential is simply another Dirac comb potential in the wavevector space: $V(\mathbf{r}) = S\sum\delta(\mathbf{r}-\mathbf{R_n}) = \sum_n \frac{S}{\Omega}e^{i\mathbf{G_n}\cdot\mathbf{r}} \implies V(\mathbf{q}) = \sum_n \mathcal{V}_{\mathbf{G}_n}\delta(\mathbf{q}-\mathbf{G}_n)$, where Ω is the volume of the BZ in real space, $\mathbf{G}'s$ are the reciprocal lattice vectors, and $\mathcal{V}_{\mathbf{G}_n}$ are the Fourier coefficients.

The defining property of the Dirac comb potential is that all Fourier coefficients $\mathcal{V}_{\mathbf{G}_n}$ are identical. From a perturbation theory perspective in the matrix eigenvalue representation, this means every electron $\mathbf{k}$-state is coupled to every other $\mathbf{k} \pm \mathbf{G}$ state with the *same strength*, and all off-diagonal matrix elements take the same value. The *exact bandstructure* for the 3D Dirac comb potential therefore can be calculated by assuming the nearly free-electron states for the diagonal elements, and the Fourier coefficient of the Dirac potential as the off-diagonal elements. Exercise 13.2 discusses the modification to the nearly free-electron bandstructure for the FCC lattice caused by a Dirac comb potential.

In real crystals, the potential $V(r)$ experienced by electrons is not the faceless, shapeless forms that the Dirac delta function models. The real periodic crystal potential varies within the unit cell, and has the geometry and the symmetries of the crystal baked in. While the 3D Dirac comb potential can capture the symmetries of the lattice, it cannot capture the details of the atomic potentials. This detail is responsible for the rich variation of properties exhibited by crystals of different chemical compositions, such as say Si, Ge and GaAs. For example, consider the potential experienced by any of the 4 valence electrons in the outermost shell of Ge or Si due to all the core electrons and the nucleus. It may be approximated by the Coulomb potential $V_{actual}(r) = -(4e^2)/4\pi\epsilon_0 r$ if all the net $+4e$ charge were precisely a point charge located at the nucleus. Fig. 13.8 shows this schematically. But because the core electrons are not located at the nucleus, and extend out to an effective radius r_c, the valence electrons experience an effective *pseudo*-potential and will be mostly unaware of the true $V(r)$ inside the core radius, but sensitive to $V(r)$ outside. Thus, to calculate the wavefunctions and energies of only the valence electrons, the exact potential and wavefunction are traded off for the smooth pseudopotential and the resulting smooth pseudo-wavefunction as shown in Fig. 13.8.

If the core electrons were all located at the nucleus, the Fourier transform of the resulting Coulomb potential would take the shape of the well-known form $V(q) = -(4e^2)/\epsilon q^2$. But because of screening, the

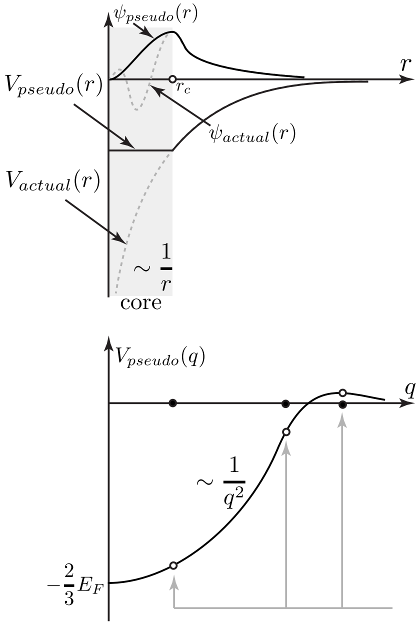


Fig. 13.8 Concept of a pseudopotential and a pseudo-wavefunction near the nucleus of an atom. The central idea is that the deep core-level electrons experience the full Coulomb potential, whereas the valence electrons only experience an effectively screened pseudopotential, which can be used to construct a smooth wavefunction free of wiggles near the nucleus.

Fig. 13.9 James Phillips, who investigated chemical bonding and bands in solids, was one of the earliest to develop the pseudopotential bandstructure to understand semiconductors. He wrote an influential book called ”Bonds and Bands in Semiconductors”.

$1/q^2$ form is smoothed out as shown in Fig. 13.8. The minimum energy is roughly of the order of the Fermi energy for this potential, which would imply that the bandwidth of electron energies in the valence band should be of the order of $E_F = (\hbar^2 k_F^2)/(2m_e)$ with $k_F = (3\pi^2 n_v)^{\frac{1}{3}}$ if the valence electrons of density n_v were nearly free. In other words, a large part of the electron energies in the valence band will be close to the nearly free–electron bands. Some of the earliest experimental data measured for metals (e.g. Al, Cu), and silicon and germanium by X-ray Photoelectron Spectroscopy (XPS) showed that the energy bandwidth of the valence band was indeed about $E_F \approx (\hbar^2 (3\pi^2 n_v)^{\frac{2}{3}})/(2m_e)$, and provided the first motivation for the pseudopotential bandstructure method. A crucial aspect is that since the atomic potential is periodic in the real space lattice, its Fourier components at the points $q_{nm} = |\mathbf{G_n} - \mathbf{G_m}|$ are the only ones that valence electron waves experience. As a consequence, the entire bandstructure is found by just knowing the values of the pseudopotential at those specific points $V_{pseudo}(q_{nm})$ as shown in Fig. 13.8. Since for large q the pseusopotential falls off fast as $1/q^2$, the number of pseudopotential values needed to calculate the entire bandstructure can be as small as 3 for Si and Ge[4]! Though modern first principle methods give a method to calculate these pseudopotentials, they are best concluded from experimental measurements of optical spectra as will be described in Chapter 27. When the experimental data is used to find the pseudopotentials, this method is referred to as the **Empirical Pseudopotential Method** (EPM) for bandstructure calculation. Texts referred to at the end of this chapter provide more rigorous justification of the use of the pseudopotential technique. We proceed directly to demonstrate its use in calculating the bandstructures for a few semiconductors.

The starting point for pseudopotential bandstructure calculation is familiar and simple: it is exactly the same nearly free-electron bandstructure of the empty lattice that was described in Chapter 10. The nearly free–electron eigenvalues are, as usual,

$$E(k_x, k_y, k_z) = \frac{\hbar^2}{2m_e}[(k_x - G_x)^2 + (k_y - G_y)^2 + (k_z - G_z)^2]. \tag{13.18}$$

By choosing a set of $\mathbf{G}$ vectors with various (l, m, n), we create the **pseudopotential matrix**

$$\begin{array}{cccccc}
 & |\mathbf{k}-\mathbf{G}_1\rangle & |\mathbf{k}-\mathbf{G}_2\rangle & |\mathbf{k}-\mathbf{G}_3\rangle & \ldots & \ldots \\
\langle\mathbf{k}-\mathbf{G}_1| & \frac{\hbar^2}{2m_e}(\mathbf{k}-\mathbf{G}_1)^2 & V_{ps}(\mathbf{G}_1-\mathbf{G}_2) & V_{ps}(\mathbf{G}_1-\mathbf{G}_3) & \ldots & \ldots \\
\langle\mathbf{k}-\mathbf{G}_2| & V_{ps}(\mathbf{G}_2-\mathbf{G}_1) & \frac{\hbar^2}{2m_e}(\mathbf{k}-\mathbf{G}_2)^2 & V_{ps}(\mathbf{G}_2-\mathbf{G}_3) & \ldots & \ldots \\
\langle\mathbf{k}-\mathbf{G}_3| & V_{ps}(\mathbf{G}_3-\mathbf{G}_1) & V_{ps}(\mathbf{G}_3-\mathbf{G}_2) & \frac{\hbar^2}{2m_e}(\mathbf{k}-\mathbf{G}_3)^2 & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots
\end{array}, \tag{13.19}$$

where the diagonal elements are the nearly free–electron energies and off-diagonal elements are pseudopotentials of the form

---

[4] One must admit that being able to conclude the energy eigenvalues and eigenfunctions of electrons over the entire Brillouin zone k-space resulting from the complex crystal potential with as few as three numbers is indeed quite remarkable.

$$V_{ps}(\mathbf{G}_i - \mathbf{G}_j) = S^S(\mathbf{G}_i - \mathbf{G}_j) \cdot V^S(|\mathbf{G}_i - \mathbf{G}_j|^2) + \\ iS^A(\mathbf{G}_i - \mathbf{G}_j) \cdot V^A(|\mathbf{G}_i - \mathbf{G}_j|^2), \tag{13.20}$$

where $S^S(\mathbf{G}_i - \mathbf{G}_j)$ and $S^A(\mathbf{G}_i - \mathbf{G}_j)$ are the symmetric and antisymmetric parts of the **structure factor** that captures the lattice symmetries, and $V^S(|\mathbf{G}_i - \mathbf{G}_j|^2)$ and $V^A(|\mathbf{G}_i - \mathbf{G}_j|^2)$ are the symmetric and antisymmetric parts of the pseudopotential of the particular crystal, which captures the chemical nature. The pseudopotential electron states are linear combinations of the plane waves $|\psi\rangle = \sum_G c_G |k + G\rangle$, and the problem is solved completely by evaluating the eigenvalues and eigenfunctions of the equation

$$\begin{pmatrix} \frac{\hbar^2}{2m_e}(\mathbf{k}-\mathbf{G}_1)^2 & V_{ps}(\mathbf{G}_1-\mathbf{G}_2) & V_{ps}(\mathbf{G}_1-\mathbf{G}_3) & ... & ... \\ V_{ps}(\mathbf{G}_2-\mathbf{G}_1) & \frac{\hbar^2}{2m_e}(\mathbf{k}-\mathbf{G}_2)^2 & V_{ps}(\mathbf{G}_2-\mathbf{G}_3) & ... & ... \\ V_{ps}(\mathbf{G}_3-\mathbf{G}_1) & V_{ps}(\mathbf{G}_3-\mathbf{G}_2) & \frac{\hbar^2}{2m_e}(\mathbf{k}-\mathbf{G}_3)^2 & ... & ... \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \end{pmatrix} \begin{pmatrix} c_{\mathbf{G}_1} \\ c_{\mathbf{G}_2} \\ c_{\mathbf{G}_3} \\ ... \\ ... \end{pmatrix} = E(\mathbf{k}) \begin{pmatrix} c_{\mathbf{G}_1} \\ c_{\mathbf{G}_2} \\ c_{\mathbf{G}_3} \\ ... \\ ... \end{pmatrix}, \tag{13.21}$$

which is solved for the eigenvalues $E(\mathbf{k})$ that form the bandstructure of the semiconductor. Each energy eigenvalue corresponds to an eigenvector $(c_{\mathbf{G}_1}, c_{\mathbf{G}_2}, ...)$ which gives the corresponding electron eigenfunction $\psi(\mathbf{r}) = \sum_n c_{\mathbf{G}_n} e^{i(\mathbf{k}+\mathbf{G}_n)\cdot\mathbf{r}}$. This eigenfunction may be plotted in real space to visualize the electron bonds, and also used to evaluate matrix elements required for transport or optical properties. The size of the matrix is determined by the number of plane waves based on the choices of $\mathbf{G}$. For example, including reciprocal lattice vectors indexed as (n_1, n_2, n_3) over the range $(-3,-3,-3)...(0,0,0)...(+3,+3,+3)$ implies 7^3 distinct values of $\mathbf{G}_n$, and therefore a 343×343 matrix[5].

13.7 Bandstructure of Si, Ge, and GaAs

A large number of semiconductors belong to the FCC lattice structure of lattice constant a. The real space lattice vectors are $\mathbf{a}_1 = \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}})$, $\mathbf{a}_2 = \frac{a}{2}(\hat{\mathbf{y}} + \hat{\mathbf{z}})$, and $\mathbf{a}_3 = \frac{a}{2}(\hat{\mathbf{z}} + \hat{\mathbf{x}})$, and the volume of the unit cell is $\Omega_r = \frac{a^3}{4}$. The corresponding reciprocal lattice vectors needed for the pseudopotential bandstructure calculations are then $\mathbf{b}_1 = \frac{2\pi}{a}(1,1,-1)$, $\mathbf{b}_2 = \frac{2\pi}{a}(1,-1,1)$, and $\mathbf{b}_3 = \frac{2\pi}{a}(-1,1,1)$, which form the general FCC reciprocal lattice vector

$$\mathbf{G} = n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3 = \frac{2\pi}{a}(n_1 + n_2 - n_3, n_1 - n_2 + n_3, -n_1 + n_2 + n_3). \tag{13.22}$$

The nearly free–electron bandstructure for the FCC lattice along the $L \to \Gamma \to X \to U, K \to \Gamma$ high symmetry path that was discussed

---

[5] The eigenvalues and eigenfunctions of a matrix of this size is obtained within seconds or minutes on a modern personal computer. Because of this ease of computation, the reader is strongly encouraged to set up the following pseudopotential bandstructure examples on his/her own computer.

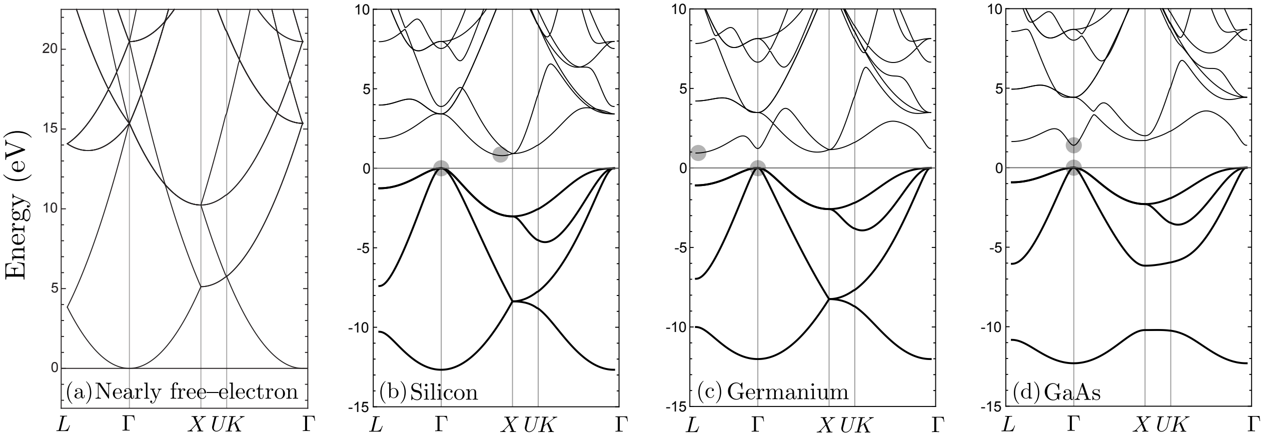


Fig. 13.10 Bandstructures of (a) nearly free–electron, and that of (b) Si, (c) Ge, and (d) GaAs calculated by the empirical pseudopotential method. Note the similarities in several areas of the semiconductor bandstructure to the nearly free–electron bandstructure far from the high symmetry points. In those sections of the bandstructure, it is as if the electron does not see the atoms! Note the energy axes *range* are the same for all the panels, though the absolute values are not.

Table 13.1 Psedopotential parameters for cubic semiconductors in $R_y = 13.6$ eV units. **Basic Semiconductor Physics** by C. Hamaguchi provides tables of such pseudoptentials including those of Table 13.2.

	Si	Ge	GaAs	
a (Å)	5.43	5.66	5.64	
V_3^S	-0.21	-0.23	-0.23	
V_8^S	+0.04	+0.01	+0.01	
V_{11}^S	+0.08	+0.06	+0.06	
V_3^A	0	0	+0.07	
V_4^A	0	0	+0.05	
V_{11}^A	0	0	+0.01	
	AlAs	InAs	GaP	InP
a (Å)	5.66	6.04	5.44	5.86
V_3^S	-0.221	-0.22	-0.22	-0.23
V_8^S	+0.025	0.00	+0.03	+0.01
V_{11}^S	+0.07	+0.05	+0.07	+0.06
V_3^A	+0.08	+0.08	+0.12	+0.07
V_4^A	+0.05	+0.05	+0.07	+0.05
V_{11}^A	-0.004	+0.03	+0.02	+0.01
	InSb	GaSb	AlSb	
a (Å)	6.48	6.12	6.13	
V_3^S	-0.20	-0.22	-0.21	
V_8^S	+0.00	+0.01	+0.02	
V_{11}^S	+0.04	+0.05	+0.06	
V_3^A	+0.06	+0.06	+0.06	
V_4^A	+0.05	+0.05	+0.04	
V_{11}^A	+0.01	+0.01	+0.02	

in Chapter 10 is reproduced in Fig. 13.10 (a) for comparison with the pseudopotential bandstructures of various semiconductors. With the vectors $\mathbf{r}_A = \frac{a}{8}(1,1,1)$ and $\mathbf{r}_B = -\frac{a}{8}(1,1,1)$, the structure factor $S^A(\mathbf{G}) = \frac{e^{i\mathbf{G}\cdot\mathbf{r}_A}}{2}$ and $S^B(\mathbf{G}) = \frac{e^{i\mathbf{G}\cdot\mathbf{r}_B}}{2}$ lead to the pseudopotential matrix elements

$$V_{ps}(\mathbf{G}_i - \mathbf{G}_j) = V^S(|\mathbf{G}_i - \mathbf{G}_j|^2)\cos(\frac{\pi}{4}(n_1 - m_1, n_2 - m_2, n_3 - m_3)) + \\ iV^A(|\mathbf{G}_i - \mathbf{G}_j|^2)\sin(\frac{\pi}{4}(n_1 - m_1, n_2 - m_2, n_3 - m_3)), \tag{13.23}$$

where (n_1, n_2, n_3) and (m_1, m_2, m_3) define the reciprocal lattice vectors. The value of $|\mathbf{G}_i - \mathbf{G}_j|^2$ ranges from 0 to 27 in units of $(2\pi/a)^2$ as the indices (n_1, n_2, n_3) and $m_1, m_2, m_3)$ are varied from $(0,0,0)$ to $(\pm 3, \pm 3, \pm 3)$. A majority of pseudopotential terms vanish due to the cos(...) and sin(...) terms in Equation 13.23; the smallest three that survive have $|\mathbf{G}|^2$ values of 3, 8, 11. The symmetric and antisymmetric pseudopotential coefficients for these surviving terms are listed in Table 13.1 in Rydberg ($= 13.6$ eV) units. The strength of the pseudopotential changes with $q^2 = |\mathbf{G}|^2$ roughly as was anticipated by the upward pointing arrows in Fig. 13.8. With Table 13.1 and Equation 13.21, the bandstructure of a large number of semiconductors of the FCC lattice structure are easily evaluated by the empirical pseudopotential method.

The resulting bandstructures of silicon, germanium, and GaAs are shown in Fig. 13.10. Thick lines are valence bands and thin lines are conduction bands. The bandstructure of Si and Ge only needs three pseudopotentials V_3^S, V_8^S, and V_{11}^S; the antisymmetric terms vanish because both atoms in the basis are identical. The indirect bandgap nature of silicon and germanium are highlighted by the conduction band minima along the $\Gamma - X$ direction for silicon and at the L-point for germanium. The valence bands of silicon and germanium appear nearly identical in shape over this energy scale. Note that for silicon, the energy separation between the lowest conduction band and the highest valence band is nearly the same along the entire **k**-path plotted: this leads to a rather large absorption peak for a photon of this energy. This optical critical point behavior will be discussed in Chapter 27. There is a X-point valence band crossing at $\sim$ 8 eV below the valence band maximum labeled at 0 eV for silicon and germanium. This degeneracy is lifted in the bandstructure of GaAs because the antisymmetric pseudopotentials are non-zero on account of the non-identical Ga and As atoms in the basis. It is highly instructive to compare the pseudopotential bandstructures of silicon, germanium, and GaAs in Fig. 13.10 with the nearly free–electron bandstructure of Fig. 13.10 (a). The lowest valence bands are nearly unchanged, except at the L point, as are some more valence bands. For semiconductors, most changes occur at the $\sim$ 15 eV Γ-point range: the pseudopotential terms split the large number of bands that cross there to open energy gaps. Some sections of the conduction bands at higher energies are also seen to be nearly unchanged from the nearly free–electron case.

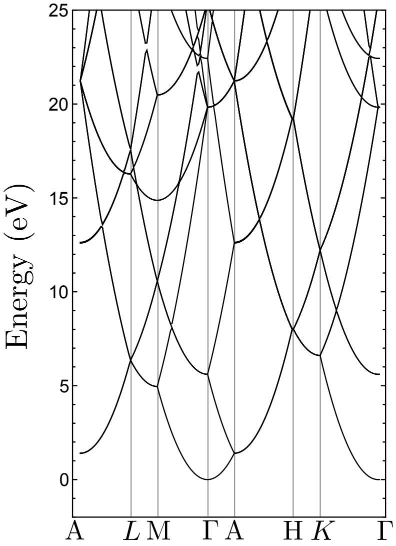


Fig. 13.11 Nearly free–electron bandstructure for a $a = 3.18$ Å, $u = 0.377$ HCP lattice.

13.8 Bandstructure of AlN, GaN, and InN

We next consider the bandstructure of the nitride semiconductors GaN, AlN, and InN, which are used for visible and ultraviolet photonics, as well as for ultrafast and high-voltage electronics. These semiconductors are of the wurtzite crystal structure, with a hexagonal close-packed (HCP) lattice. The nearly free–electron bandstructure for the HCP lattice shown in Fig. 13.11 was discussed in Chapter 10 (Fig. 10.12). The real space lattice vectors are $\mathbf{a}_1 = a\hat{\mathbf{x}}$, $\mathbf{a}_2 = a(\frac{1}{2}\hat{\mathbf{x}} + \frac{\sqrt{3}}{2}\hat{\mathbf{y}})$, and $\mathbf{a}_3 = c\hat{\mathbf{z}}$, and the volume of the unit cell is $\Omega_r = \frac{\sqrt{3}}{2}a^2c$. The corresponding reciprocal space lattice vectors are $\mathbf{b}_1 = \frac{2\pi}{a}(\hat{\mathbf{x}} - \frac{1}{\sqrt{3}}\hat{\mathbf{y}})$, $\mathbf{b}_2 = \frac{2\pi}{a}\frac{2}{\sqrt{3}}\hat{\mathbf{y}}$, and $\mathbf{b}_3 = \frac{2\pi}{c}\hat{\mathbf{z}}$, with unit cell volume $(2\pi)^3/\Omega_r$. The reciprocal lattice vectors are then given by $\mathbf{G} = m_1\mathbf{b}_1 + m_2\mathbf{b}_2 + m_3\mathbf{b}_3$. Defining $c = a/\sqrt{u}$, the dimensionless ratio $u = (\frac{a}{c})^2$ is used to characterize the semiconductor. For an ideal HCP structure, $\frac{c}{a} = \sqrt{\frac{8}{3}}$, or $u = \frac{3}{8}$. The equilibrium lattice constants of GaN, AlN, and InN differ from this ideal value. The square of the reciprocal lattice vector of indices (l, m, n) then is

$$|\mathbf{G}|^2 = (\frac{2\pi}{a})^2[l^2 + \frac{1}{3}(2m - l)^2 + un^2]. \tag{13.24}$$

The pseudopotential matrix elements are

$$V_{ps}(\mathbf{G}_i - \mathbf{G}_j) = S^S(\mathbf{G}_i - \mathbf{G}_j) \cdot V^S(|\mathbf{G}_i - \mathbf{G}_j|^2) + \\ iS^A(\mathbf{G}_i - \mathbf{G}_j) \cdot V^A(|\mathbf{G}_i - \mathbf{G}_j|^2), \tag{13.25}$$

Table 13.2 Psedopotential parameters for a few wurtzite semiconductors.

	$V^S(q)$	$V^A(q)$
GaN		
a_1	0.04258	0.5114
a_2	13.079	−20.122
a_3	0.226	0.0835
a_4	20.393	−41.557
AlN		
a_1	0.0363	0.0323
a_2	11.960	−145.212
a_3	0.234	0.0947
a_4	22.233	−19.160
InN		
a_1	0.0459	0.0221
a_2	12.542	−35.605
a_3	0.299	0.0574
a_4	17.691	−18.261

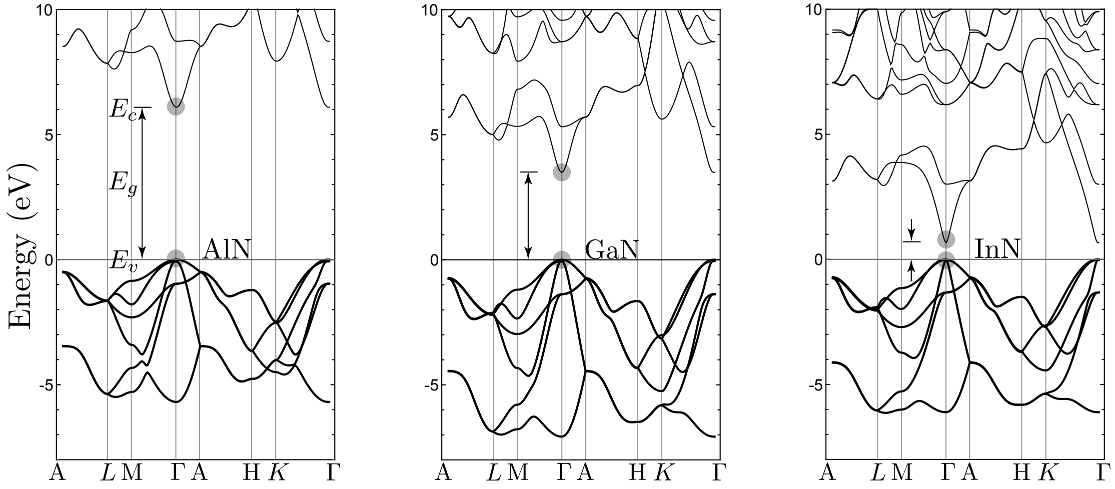


Fig. 13.12 Bandstructures of wurtzite AlN, GaN, and InN calculated by the empirical pseudopotential method.

where the HCP structure factors are

$$S^S = \cos[2\pi(\frac{l}{6}+\frac{m}{6}+\frac{n}{4})]\cos[\frac{2\pi nu}{2}]$$
$$S^A = \cos[2\pi(\frac{l}{6}+\frac{m}{6}+\frac{n}{4})]\sin[\frac{2\pi nu}{2}], \tag{13.26}$$

and the pseudopotential is expressed as a function of wavevector q:

$$V(q) = \frac{a_1(2q^2-a_2)}{1+e^{a_3(2q^2-a_4)}}, \tag{13.27}$$

where the a_i values are given in Table 13.2 individually for the symmetric and the asymmetric terms of the pseudopotential and plotted in Fig. 13.14. Note that pseudopotentials are retained for $|\mathbf{G}|^2 < 16$.

The resulting bandstructures for AlN, GaN, and InN are shown in Fig. 13.12. All three nitrides are direct-bandap semiconductors. In addition, if the conduction bands are considered alone, the intervalley separation between the lowest Γ-point energy and the next lowest conduction band valleys are nearly 1.5 eV or more higher in energy; the comparative intervalley energy separation is quite a bit smaller in GaAs, Ge, and somewhat small in silicon. Compared to the cubic crystals of Fig. 13.10, the valence band widths are observed to be narrower for the nitride semiconductors. The highest (heavy-hole) valence band has far less curvature than the conduction band, highlighting the large heavy hole effective mass in the nitrides. It is again instructive to compare all three bandstructures of Fig. 13.12 with the nearly free–electron bandstructure of the HCP lattice in Fig. 13.11.

The bandgaps and critical points in optical spectra of the nitride family of semiconductors are captured well by the empirical pseudopotential bandstructures. Fig. 13.13 shows the crystal structure

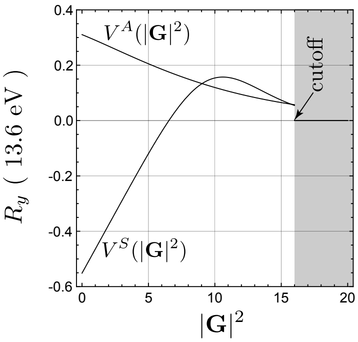


Fig. 13.14 Pseudopotential Matrix Elements for GaN, expressed as a continuous function of $|\mathbf{G}_1 - \mathbf{G}_2|^2$.

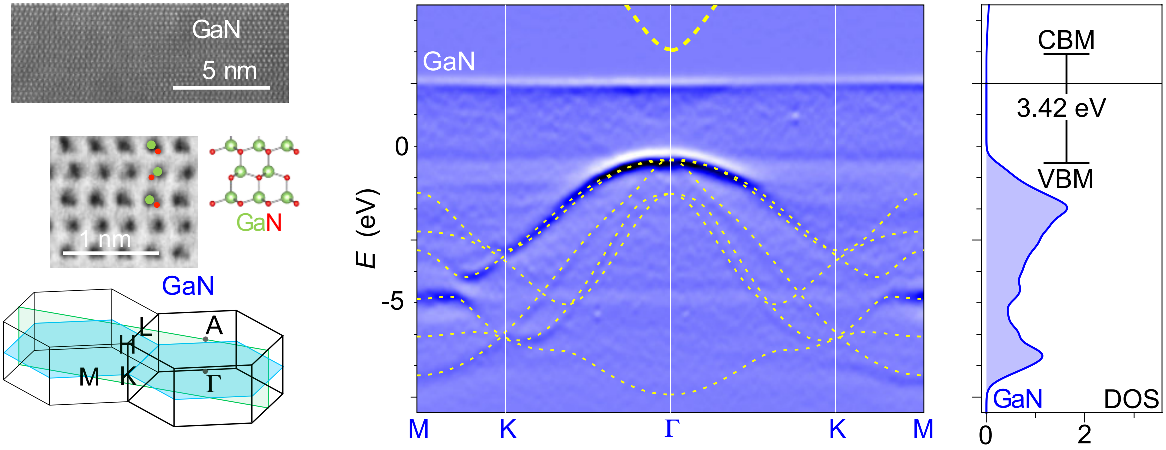


Fig. 13.13 Experimentally measured valence bandstructure of GaN compared to the DFT calculation. The DFT calculation is nearly identical to the pseudopotential bandstructure shown in Fig. 13.12. The experimental measurement uses angle-resolved photoelectron spectroscopy (ARPES). Courtesy of V. Strocov, T. Yu, J. Wright, C. Chang, B. Pamuk, and G. Khalsa.

of GaN indicating the individual atoms arranged in the HCP lattice as imaged by transmission electron microscopy. The Brillouin zone is shown with the high symmetry points. A color figure of the experimentally measured valence band structure of GaN, and the density functional theory (DFT) calculated dashed lines are shown, along with the density of states of the valence band. Comparison of the empirical pseudopotential bandstructure and the DFT bandstructure indicate they are nearly identical for the valence band. The difference in the methods is the pseudopotential given by Equation 13.27 are empirically determined from comparison to experiment, whereas that of DFT is calculated from first principles. We give a brief discussion of the DFT method to wrap up the discussion on electron bandstructures of semiconductors.

13.9 Pseudopotentials to DFT and beyond

Fermi and Thomas had introduced pseudopotentials as a means to approximate the distribution of valence electrons in atoms. The valence electrons are screened from the Coulomb potential of the nucleus and core electrons (this is called Thomas–Fermi screening), and the resulting Fermi energy becomes a function of the volume density of the valence electrons, i.e. $E_F = f(n_v)$. The pseudopotential methods, as well as DFT is traced back to this idea. Kohn and Hohenberg proved an important theorem that the ground-state energy of a quantum mechanical system governed by the Schrödinger equation is a function of the density function $n(r)$ of electrons, which varies in space as the square of the wavefunction $n(r) \sim |\psi(r)|^2$. In other words, to find the ground state energy we can get away without knowing the wavefunction, as long as we know the density. Because the energy is a function of the density function, it is a functional[6]. In DFT, instead of determining the potential experienced by electrons by empirical methods, it is calculated *ab initio* by solving the Schrödinger equation using the Kohn–Hohenberg theorem.

This method is computationally more intensive and therefore the number of atoms that can be simulated is restricted. Difficulties in obtaining accurate bandgaps of semiconductors have been addressed recently by the introduction of many-body corrections and hybrid functionals. The real power of the DFT methods are the host of physical properties that can be computed using it beyond the bandstructure, such as the total energy of the crystal, which is minimized at the equilibrium lattice constants and crystal structure, and the energetic preference of defects and dopants in the semiconductor crystals. Because bandstructures desired to simulate the operation of electronic and photonic devices need to be computationally efficient, often DFT is used to find accurate parameters for either pseudopotential, tight-binding, or $\mathbf{k} \cdot \mathbf{p}$ bandstructures. Once the parameters are determined, these techniques are far more nimble than DFT for device level simulations.

This section marks the end of the extensive discussion of the computation of semiconductor bandstructures in the past few chapters. In the following chapters extensive use of the properties discussed will be made, both for bulk semiconductors and their quantized heterostructures. The recipe to capture the bands of doped semiconductors and heterostructures is the subject of the next Chapter (14). Before you leave this chapter, you will benefit by circling back to the discussion in Chapter 9, Section 9.2 on how the Bloch theorem led to all four recipes of bandstructure: (1) the nearly free–electron bandstructure (Chapter 10), (2) tight-binding bandstructure (Chapter 11), (3) $\mathbf{k} \cdot \mathbf{p}$ bandstructure (Chapter 12), and (4) exact and empirical pseudopotential bandstructure (this Chapter). The quest for a more perfect experimental measurement and computation and understanding of bandstructures continues to be at the forefront of semiconductor physics in particular, and condensed matter physics in general.

13.10 Chapter summary section

In this chapter, we learned:

- The exact bandstructure for Dirac comb potentials, and its use to model defects,
- The empirical pseudopotential method of bandstructure calculations, and
- The hierarchy of various bandstructure calculation methods.

[6] A functional is a function of a function.

Further reading

Electronic Structure and Optical Properties of Semiconductors by M. L. Cohen and J. R. Chelikowsky, written by the original architects of the pseudopotential method for semiconductors, remains the gold standard for the subject. **Fundamentals of Semiconductors** by P. Yu and M. Cardona has excellent discussion of bandstructure calculation methods in general, and the empirical pseudopotential method in particular. **Basic Semiconductor Physics** by C. Hamaguchi is a good resource for several pseudopotential bandstructure calculation details.

Exercises

(13.1) **Defect states in the Kronig–Penney model**

We have discussed about the Kronig–Penney crystal model, and how defect states cause energy eigenvalues to split off from band-edges. For simplicity, we are going to solve this problem for $k = 0$ (the Γ point), and neglect all other k-points. Let us say that a perfect Kronig–Penney crystal of lattice constant a has N eigenvalues E_i with $i = 1...N$ in a band at Γ. All other energy bands are very far in energy, and may be neglected. If at one of the lattice sites, a defect changes the delta function strength by U_0, then the new *exact* eigenvalues are given by solving

$$\frac{a}{U_0} = \text{Trace}[G(E)] = \sum_{i=1}^{N} \frac{1}{E - E_i}, \tag{13.28}$$

for allowed energies E, where U_0 is in eV-nm units, $G(E)$ is the Green's function, the Trace of which is just the inverse sum on the right.

(a) Argue why the equation above is correct if no defect is present.

(b) Show *graphically* that if the eigenvalues themselves are widely separated, and $U_0 >> 0$, the eigenvalue that splits off the band due to the defect has energy $E_s^+ \approx E_N + \frac{U_0}{a}$.

(c) Using your graph of part (b), show that if $U_0 << 0$, the eigenvalue that splits off the band due to the defect has energy $E_s^- \approx E_1 + \frac{U_0}{a}$.

(13.2) **Bandstructure of a FCC lattice with a Dirac comb potential**

Fig. 13.15 shows the calculated bandstructure for a FCC lattice of lattice constant $a = 5.43$ Å for the nearly free–electron model as the dashed lines, and a periodic Dirac comb potential with $V(\mathbf{q}) = -0.2$ eV for all $\mathbf{q} = \mathbf{G_n} - \mathbf{G_m}$.

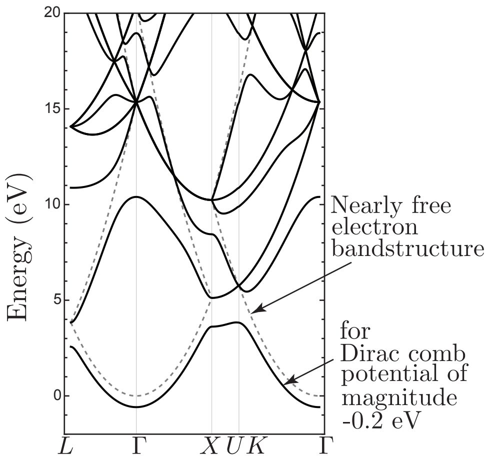


Fig. 13.15 Bandstructure for a weak Dirac comb potential of the form $V(\mathbf{r}) = S\sum\delta(\mathbf{r} - \mathbf{R_n})$ for a FCC lattice, compared to the nearly free–electron bandstructure for which $V(r) = 0$.

(a) Explain why the periodic potential moves the eigenvalues in the lowest band in the manner seen in the figure.

(b) Note that some sections of bands are nearly unaffected by the periodic potential. Explain why this is the case.

(c) Calculate and reproduce this figure, and repeat for a Dirac comb potential of strength +0.2 eV instead of the negative potential used here.

(13.3) **Pseudopotential bandstructures of various direct and indirect bandgap semiconductors**
Calculate using the pseudopotential method, the bandstructures of various semiconductors for which the parameters and lattice constants are provided in the chapter. Compare with the bandstructures shown in the chapter and check if the calculated bandstructures capture the known experimental bandgaps and whether they are direct or indirect.

(13.4) **Branch points and band-alignments**

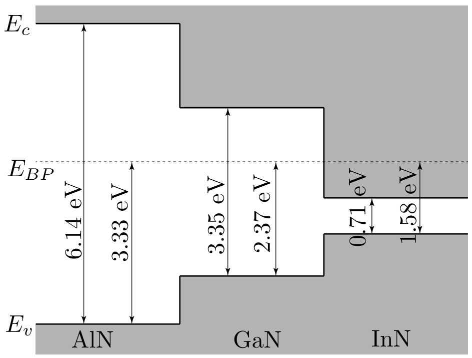


Fig. 13.16 Bandstructure for a weak Dirac comb potential of the form $V(\mathbf{r}) = S\sum\delta(\mathbf{r}-\mathbf{R_n})$ for a FCC lattice, compared to the nearly free–electron bandstructure for which $V(r) = 0$.

This Exercise follows the paper Applied Physics Letters, vol. 94, page 012104 (2009). Read this paper and answer the following questions. There exists an unique energy value E_{BP} for any bandstructure, for which we have the condition

$$\sum_{\mathbf{k}}[E_{BP} - E_{lower}(\mathbf{k})] = \sum_{\mathbf{k}}[E_{higher}(\mathbf{k}) - E_{BP}]. \tag{13.29}$$

This branch point energy is, in a sense, the "weighted average" energy level. Eigenvalues below this energy are valence-band like – meaning they tend to be occupied and are ready to donate electrons, and therefore are donor-like. Eigenvalues above this energy are conduction-band like, meaning they are empty and can accept electrons and therefore are acceptor-like. E_{BP} is therefore also related to the propensity of charge transfer between semiconductors at interfaces, and is also called the **charge neutrality level**.

(a) For n_+ bands with higher energy than E_{BP} and n_- bands with lower energy using a sum over $\mathbf{k}$ runs over the entire Brillouin zone, say over N states, which is equal to the macroscopic number of lattice points in the crystal show that

$$N \cdot n_- \cdot E_{BP} - \sum_{\mathbf{k},i} E_{lower,i}(\mathbf{k}) = \sum_{\mathbf{k},j} E_{higher,j}(\mathbf{k}) - N \cdot n_+ \cdot E_{BP}, \tag{13.30}$$

which gives the Branch point energy at

$$E_{BP} = \frac{\sum_{\mathbf{k},i} E_{lower,i}(\mathbf{k}) + \sum_{\mathbf{k},j} E_{higher,j}(\mathbf{k})}{N(n_- + n_+)}. \tag{13.31}$$

(b) Apply this result to calculate the branch point of silicon using the pseudopotential bandstructure by retaining the top three valence bands and the lowest conduction band. How close is E_{BP} to the mid-gap state?

(c) Argue why for indirect bandgap semiconductors such as silicon, E_{BP} is close to the middle of the bandgap.

(d) Show why the branch point energy of direct-gap semiconductors is

$$E_{BP} = \frac{1}{2N_k}\sum_k \left[\frac{\sum_i^{N_{CB}} E_{ci}(\mathbf{k})}{N_{CB}} + \frac{\sum_j^{N_{VB}} E_{vj}(\mathbf{k})}{N_{VB}}\right], \tag{13.32}$$

where N_k is the number of points in the k-space chosen for the sum, N_{CB} and N_{VB} are the conduction and valence band points of eigenvalues $E_{ci}(\mathbf{k})$ and $E_{vj}(\mathbf{k})$ respectively.

(e) As discussed in the paper, E_{BP} in direct-bandgap semiconductors can deviate from the mid-gap states. Apply the results to find the E_{BP} of AlN, GaN, and InN using their pseudopotential bandstructures. Compare with Fig. 13.16. Note that if the branch-points align, the band-offsets are known immediately if the bandgaps are known.

(f) Note that unlike AlN and GaN, the branch-point of InN is deep inside the conduction band. What does this mean for the type of conductivity of the semiconductor?

(13.5) **Density of states from pseudopotential bandstructure**
The density of states (DOS) is obtained from the calculated bandstructure. Using the pseudopotential bandstructures of silicon, GaAs, and GaN, obtain the DOS $g(E)$ and plot them as $g(E)$ vs. E using the following methods.

(a) Identify the Brillouin zone and create a uniform mesh of **k**-points. Calculate the eigenvalues at each **k**-point, and place the values in discrete bins of energies $\Delta E = E_{i+1} - E_i$. Make a plot of this histogram – this is the DOS. Compare with the DOS shown in Fig. 11.18.

(b) Argue why the brute-force method of part (a) is very wasteful of computational resources, and may miss important points in the Brillouin zone too. Suggest ways to improve this procedure.

(c) A common procedure is 1) to discretize instead, the smallest irreducible portion of the Brillouin zone as determined by the symmetry, and 2) to discretize with a finer mesh around points of conduction band minima or valence band maxima, or other critical points. Modify your method and evaluate the DOS using this scheme and compare the speed and accuracy of the DOS with those obtained by the brute force method in part (a).

(13.6) **Electron densities from pseudopotential bandstructure**
In the chapter we discussed primarily the energy eigenvalues $E_n(\mathbf{k})$ of bands in semiconductors calculated by the pseudopotential method. The same matrix solution also gives the eigenfunctions as the row vectors $[a_1, a_2, ...]$. The corresponding eigenfunctions for plane-wave bases are $\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} a_{\mathbf{G}} e^{i(\mathbf{G}\cdot\mathbf{r})}$.

(a) Calculate the eigenfunctions of the conduction band minima and valence band maxima states at the Γ-point ($\mathbf{k} = (0,0,0)$) of GaAs. Plot $|\psi_\Gamma(\mathbf{r})|^2$ and examine their symmetries with respect to the crystal directions.

(b) Compare the differences of the valence band and conduction band electron distributions $|\psi_\Gamma(\mathbf{r})|^2$. How do they vary for **k** far from the Γ-point, and near the edges of the Brillouin zone?

(c) Calculate $|\psi_\Gamma(\mathbf{r})|^2$ for the conduction band minima of silicon and compare the properties with the conduction band minima of GaAs.

(13.7) **Optical properties from pseudopotential bandstructure**
In Chapter 27 we will study the interband optical transitions of semiconductors. In the interaction of an electron in the state $\psi^v_{\mathbf{k}}(\mathbf{r})$ in the valence band of a semiconductor with light, a transition to a state $\psi^c_{\mathbf{k}}(\mathbf{r})$ in the conduction band with the same **k** is possible if a dipole matrix element of the type $\int d^3r (\psi^c_{\mathbf{k}}(\mathbf{r}))^\star \mathbf{r} \psi^v_{\mathbf{k}}(\mathbf{r})$ is non-zero for the valence band maximum and the conduction band minimum electron states.

(a) Evaluate this matrix element for silicon at the Γ-point and show that it is zero.

(b) Evaluate it for GaAs at the Γ-point and show that it is non-zero, and a direct absorption is thus allowed.

(c) Using the pseudopotential bandstructure, evaluate the matrix element for various **k** values away from the Γ-point. Combining these leads to the absorption spectrum as described in Chapter 27.

(13.8) **Pseudopotential bandstructure of aluminum and copper**
Pseudopotentials were developed in early days by noticing that for alkali metals, the experimental spread in the electron energies in photoemission measurments were almost identical to what would be expected from calculating the Fermi energy E_F from the nearly free-electron model for electrons. This in turn meant that the crystal potential was very weak and its impact on the bandstructure should be possible to model with a minimal set of Fourier weights $V_{\mathbf{G}}$ at $\mathbf{k} = \mathbf{G}$ points.

(a) Obtain from the literature the pseudopotential weights of aluminum, and using its crystal structure and lattice points, plot the bandstructure.

(b) Obtain from the literature the pseudopotential weights of copper, and using its crystal structure and lattice points, plot the bandstructure.

(c) Plot the density of states of the two metals and compare their Fermi surfaces.

(13.9) **Density functional theory**

Read the article Reviews of Modern Physics, vol. 71, page 1253 (1999) by Walter Kohn (Fig. 14.15) on the origins and development of density functional theory. Answer the following questions.

(a) Appreciate the roles of Thomas-Fermi approximation for atoms, and its role in the development of DFT. We will encounter this approximation in Chapter 23 in the model of screening by free carriers in semiconductors and their effect on electron transport.

(b) What are the computational issues encountered in applying DFT to calculate the exact bandstructure of crystals from the Schrödinger equation?

(c) Research and comment on what is known as the **bandgap problem** in DFT: that by standard functionals, the bandgaps of semiconductors are typically underestimated. Use of hybrid functionals and higher order corrections fixes this issue.

(d) Once an accurate DFT bandstructure is obtained, the empirical pseudopotentials can be extracted and the subset of numbers is useful to reproduce the entire bandstructure with high accuracy. This method is also used to find tight-binding parameters. Comment on why this parametrization of the bandstructure is useful for the computation of device level properties, when DFT becomes prohibitive.

Doping and Heterostructures: The Effective Mass Method

14

14.1 Effective mass approximation, envelope functions 292
14.2 3D, 2D, 1D, 0D: heterostructures 295
14.3 3D bulk bandstructure 296
14.4 Doped semiconductors 298
14.5 2D quantum wells 299
14.6 1D quantum wires 301
14.7 0D quantum dots 302
14.8 Finite barrier heights 304
14.9 Multilayers and superlattices 304
14.10 Wannier functions 306
14.11 Chapter summary section 307
Further reading 308
Exercises 308

The materials in electronic and photonic devices are not homogeneous and intrinsic semiconductors. Instead, they are either doped intentionally with donor or acceptor impurities, or are quantized heterostructures in the form of quantum wells, wires, or dots. Electrons and holes are typically desired not extended over the whole crystal, but in specified locations, such as inside quantum wells in high-speed transistors, or in quantum wells, or quantum dots for LEDs and lasers. The Bloch functions $\psi_{\mathbf{k}}(\mathbf{r})$ and the corresponding eigenvalues $E_n(\mathbf{k})$ form the underlying electronic basis states, but are not enough to represent electronic states that are confined near dopants, or in low-dimensional structures such as inside quantum wells or dots. This is, of course because the probability density of an electronic state represented by a Bloch function $|\psi_{\mathbf{k}}(\mathbf{r})|^2$ is the same in every unit cell of the crystal, it is *periodic*. To represent electron or hole states that are confined in real space, one must find linear combinations of the Bloch states with a collection of coefficients that represent the quantum confinement of the states appropriately. This *effective mass method* was first developed by Slater (Fig. 11.1), and by Kohn (see Fig. 14.15) and Luttinger. It is a very empowering method, since it makes the design of semiconductor quantum structures both accessible, and intuitive. The goals of this chapter are to learn:

- What is the effective mass method, and how does it map the problem of electrons in doped semiconductors or quantum heterostructures to exactly solved single-particle problems in quantum mechanics?
- Why is it a powerful tool for the *design* of doping, and quantum wells, wires and dots in semiconductors?
- How can we evaluate the density of states, carrier concentrations, and physically visualize the location of electrons or holes in confined geometries, using the idea of envelope wavefunctions?

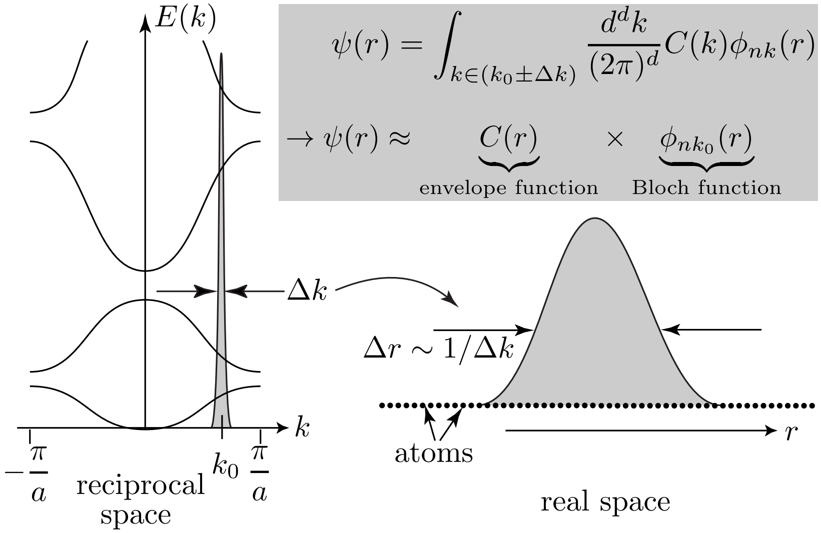

$E(k)$
$\psi(r) = \int_{k\in(k_0\pm\Delta k)} \frac{d^dk}{(2\pi)^d} C(k)\phi_{nk}(r)$
$\rightarrow \psi(r) \approx \underbrace{C(r)}_{\text{envelope function}} \times \underbrace{\phi_{nk_0}(r)}_{\text{Bloch function}}$
Δk
$\Delta r \sim 1/\Delta k$
k
$-\frac{\pi}{a}$
k_0
$\frac{\pi}{a}$
reciprocal space
atoms
r
real space

Fig. 14.1 A wavepacket is constructed by taking Bloch functions from a small region of the reciprocal space, and summing them with weights. The weights $C(k)$ have a small extent Δk in reciprocal space; when carried over to real space, the spread is large, since $\Delta r \sim 1/\Delta k$; thus the wavepacket has a finite spread in real space, and represents the wavefunction of a particle. If we restrict the sum in reciprocal space to 1% of the BZ, the wavepacket spreads over 1/0.01 = 100 atoms in real space. The real space wavefunction is given by the Bloch wavefunction at the k_0 point, modulated by an envelope function $C(r)$, which is the Fourier transform of the weights $C(k)$.

14.1 Effective mass approximation, envelope functions

So long as the perturbations of the crystal potential are not drastic, one can re-cast the Schrödinger equation in a form that is very useful for discussing transport and device applications. The uncertainty principle forbids simultaneously defining the location of a particle in real space, and it's momentum. To get around this restriction, the concept of a wave packet is necessary. Wave packets, unlike pure Bloch-eigenstates, have a finite spread both in the momentum and real space. A wave packet is a linear combination of Bloch eigenstates for small k-values around a region of interest in the Brillouin zone. For most cases, it suffices to investigate properties of electrons and holes located very close to the band extrema in the k-space; therefore, one collects Bloch eigenstates around such points, and creates a wavepacket by taking their linear combinations.

To illustrate this, let us consider the 1D case. Let us construct a *wavepacket* by taking a linear combination of the Bloch eigenstates $\phi_{nk}(x)$ from the n^{th} band with wavevector k. The sum is over the whole Brillouin zone and all bands:

$$\psi(x) = \sum_n \sum_k C(k)\phi_{nk}(x) = \sum_n \int \frac{dk}{2\pi} C(k)\phi_{nk}(x) \tag{14.1}$$

We now make two crucial approximations :

a) We assume that wavefunctions from only one band play a part in the wavepacket, and thus drop the sum over all bands.

b) We assume that in the single band we are interested in, wavevectors from a small region say around $k_0 = 0$ are important (see Fig. 14.1).

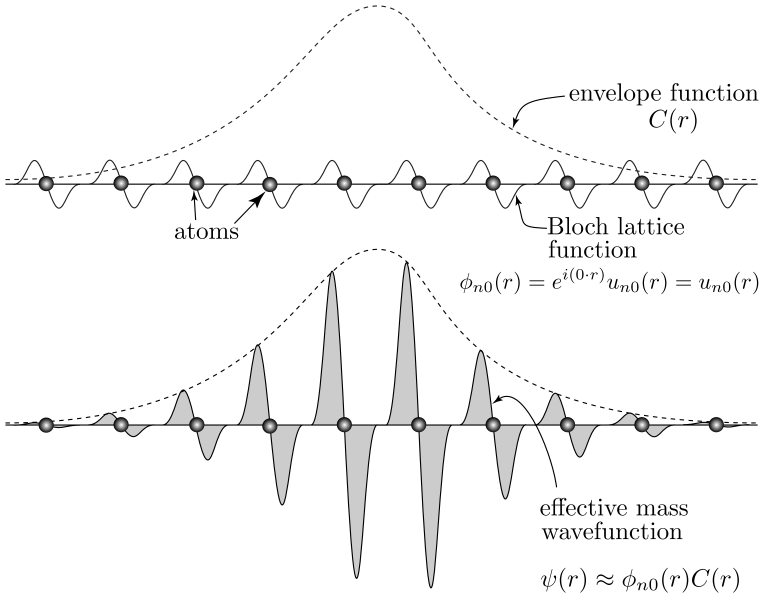

envelope function
$C(r)$
atoms
Bloch lattice function
$\phi_{n0}(r) = e^{i(0\cdot r)}u_{n0}(r) = u_{n0}(r)$
effective mass wavefunction
$\psi(r) \approx \phi_{n0}(r)C(r)$

Fig. 14.2 Envelope function $C(r)$ modulates the Bloch function $\phi_{n0}(x)$ to produce the wavefunction of the wavepacket $\psi(x)$. The total wavefunction looks similar to a slowly varying part of the form e^{ikr} modulated by a periodic part $u(r)$

Then, the Bloch functions can be written as

$$\phi_{nk}(x) = e^{ikx}u_{nk}(x) \approx u_{n0}(x)e^{ikx} = \phi_{n0}(x)e^{ikx}, \tag{14.2}$$

and the wavepacket takes the form

$$\psi(x) \approx \phi_{n0}(x) \int \frac{dk}{2\pi} C(k)e^{ikx} = \underbrace{\phi_{n0}(x)}_{\text{Bloch}} \cdot \underbrace{C(x)}_{\text{envelope}}, \tag{14.3}$$

where the integral term is identified as the Fourier transform of the weights $C(k) \leftrightarrow C(x)$. The real-space function $C(x)$, which is a Fourier transform of the weights of the wavepacket, is called the **envelope function**. Since the weights $C(k)$ are taken from over a small region in k-space, $C(x)$ is spread out over real space. It is typically a smooth function spreading over several lattice constants. This is illustrated in Figs. 14.1 and 14.2.

How does the wavepacket behave when the periodic crystal Hamiltonian $\hat{H}_0$ operator acts on it? Since $\phi_{nk}(x)$ are Bloch-eigenfunctions of this Hamiltonian, $\hat{H}_0\phi_{nk}(x) = E_n(k)\phi_{nk}(x)$, and we recover

$$\hat{H}_0\psi(x) = \int \frac{dk}{2\pi} C(k)E_n(k)\phi_{nk}(x) \approx \phi_{n0}(x) \int \frac{dk}{2\pi} C(k)E_n(k)e^{ikx}. \tag{14.4}$$

We now write out the energy eigenvalues from the known bandstructure $E_n(k)$ as a Taylor-series of small wavevectors around $k = k_0 = 0$,

$$E_n(k) = \sum_m a_m k^m, \tag{14.5}$$

and the Schrödinger equation becomes

$$\hat{H}_0\psi(x) \approx \phi_{n0}(x)\sum_m \int \frac{dk}{2\pi}C(k)a_m k^m e^{ikx}. \tag{14.6}$$

We now use a property of Fourier transforms – if $f(k) \leftrightarrow f(x)$, then $kf(k) \leftrightarrow (-id/dx)f(x)$, and in general, $k^m f(k) \leftrightarrow (-id/dx)^m f(x)$. Thus,

$$\int \frac{dk}{2\pi}k^m C(k)e^{ikx} \leftrightarrow (-i\frac{d}{dx})^m C(x), \tag{14.7}$$

and the Schrödinger equation is recast as

$$\hat{H}_0\psi(x) \approx \phi_{n0}(x)E_n(-i\nabla)C(x), \tag{14.8}$$

which can be generalized to the 3D case. Thus, in the energy term, we make the substitution $k \to i\partial/\partial r$, making it an operator that acts only on the envelope function. This step is crucial – the Bloch function part has been pulled out as a coefficient; no operators act on it. Now, instead of the periodic potential Hamiltonian, if we have another potential (say a perturbation) $W(r)$ present, the Schrödinger equation becomes

$$\hat{H}_0\phi_{n0}(r)C(r) + W(r)\phi_{n0}(r)C(r) = E\phi_{n0}(r)C(r), \tag{14.9}$$

and using Equation 14.8, it becomes

$$\boxed{[E_n(-i\nabla) + W(r)]C(r) = EC(r),} \tag{14.10}$$

where the Bloch functions do not appear at all! Furthermore, if we already know the bandstructure of the semiconductor, then we can write the energy around the point $k_0 = 0$ of interest in terms of the effective mass, and the operator $E_n(-i\nabla)$ thus becomes

$$E_n(k) \approx E_c(r) + \frac{\hbar^2 k^2}{2m^\star} \to E_n(-i\nabla) \approx E_c(r) - \frac{\hbar^2}{2m^\star}\nabla^2, \tag{14.11}$$

and the Schrödinger equation takes the enormously simplified form

$$\boxed{[-\frac{\hbar^2}{2m^\star}\nabla^2 + W(r)]C(r) = [E - E_c(r)]C(r),} \tag{14.12}$$

which is the celebrated "**effective mass approximation**". Take a moment to note what has been achieved. All information about the bandstructure and the periodic crystal potential has "vanished" into the effective mass $m^\star$. **The Schrödinger equation has been re-cast into a much simpler problem of a particle of mass $m^\star$, moving in a potential** $E_c(r) + W(r)$! The wavefunctions are envelope functions $C(r)$, from which one recovers the real wavefunction of the wavepacket by multiplying with the Bloch function – $\psi(r) \approx \phi_{n0}(r)C(r) = u_{n0}(r)C(r)$. The envelope functions $C(r)$ can be actually determined for any potential – it amounts to solving the Schrödinger equation for a particle in the potential $E_c(r) + W(r)$. Note that the envelope function in the absence of any impurity potential $W(r) = 0$ for slowly varying $E_c(r)$ is given by

$$C(r) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{14.13}$$

and the corresponding eigenvalues of the Schrödinger equation are given by

$$E = E_c(r) + \frac{\hbar^2|\mathbf{k}|^2}{2m^\star}. \tag{14.14}$$

If we consider electrons at the bottom of the conduction band, $E_c(r)$ is the spatial variation of the conduction band edge – exactly what one draws in band diagrams. An impurity potential can now be included as a perturbation to the periodic crystal, and the new energy eigenvalues can be found. As an example, consider an ionized impurity, which has a Coulomb potential. The effective mass equation takes the form

Fig. 14.3 Nevill Mott was an early pioneer of applying quantum mechanics to understand crystals, and semiconductors in particular. Mott was awarded the Nobel Prize in Physics in 1977 for his work on properties of disordered solids.

$$[-\frac{\hbar^2}{2m^\star}\nabla^2 - \frac{q^2}{4\pi\epsilon r}]C(r) = (E - E_c)C(r), \tag{14.15}$$

which is identified as the same as the classic problem of a hydrogen atom (Chapter 5, Section 5.9), albeit with two modifications – the mass term $m^\star$ is an effective mass instead of the free–electron mass m_e, and the dielectric constant ϵ_r is the permittivity inside the semiconductor instead of ϵ_0, the permittivity of vacuum. Then, the new energy levels that appear are given by

$$E - E_c = E_\infty \frac{m^\star}{\epsilon_r^2}, \tag{14.16}$$

and the effective Bohr-radius is given by

$$a_B^\star = a_B^0 \frac{\epsilon_r}{m^\star}. \tag{14.17}$$

For a typical semiconductor, $\epsilon_r \sim 10\epsilon_0$ and $m^\star \sim 0.1m_e$, which implies the effective Bohr radius is $a_B^\star \sim (10/0.1)a_B^0 \sim 100a_B^0$. Since this is nearly 100 times larger than the Bohr radius, Mott (Fig. 14.3) referred to this situation as a **swollen hydrogen atom**. The corresponding ionization energy is $\sim (10/(0.1)^2) \sim 1000\times$ smaller than the ionization energy of the hydrogen atom, or in other words, of the order of 13.6 meV instead of 13.6 eV. Since room temperature $k_bT \sim 26$ meV, the loosely bound electron is easily ionized into the band and leads to electrical conductivity. This is called n-type doping with donors.

14.2 3D, 2D, 1D, 0D: heterostructures

In bulk semiconductors, the band-edge variation in real space can be varied by applying electric fields, or by doping variations. In semiconductor heterostructures, one can further engineer the variation of the band-edge $E_c(r)$ in space by quasi-electric fields – the band edge can behave as quantum-wells, wires, or dots, depending on the composition of the semiconductor. The effective mass approximation is a natural point of departure, where analysis of such low-dimensional structures begins.

With the explosion of usage of semiconductor heterostructures in a variety of applications, low-dimensional structures such a quantum wells, wires, and dots have become important. They come in various avatars – these structures can be grown by compositional variations in epitaxially grown semiconductor layers by MBE/MOCVD techniques, or nanowires / nanotubes / nanocrystals can be grown by bottom-up approaches (by CVD techniques, or by solution chemistry). So, understanding the bandstructure of these artificially engineered materials is of great interest.

The goal of many clever expitaxial/bottom-up techniques to create nanostructures amounts to modifying the bandstructure of the constituent bulk semiconductor material. Many of these designer materials have niche applications, and have a potential to perform functions that are difficult, if not impossible to achieve in bulk materials[1].

We have derived the effective mass equation for carriers in bulk semiconductors in the envelope-function approximation. The three-dimensional effective mass equation for the conduction band is

$$[-\frac{\hbar^2}{2m^\star}\nabla^2 + V(r)]C(r) = (E - E_c(r))C(r). \qquad (14.18)$$

Here, $C(r)$ is the envelope function of carriers in the band under consideration. The Schrödinger equation is thus re-cast in a form that is identical to that of an electron in a total potential $V(r) + E_c(r)$, determined by the band-edge behavior. It has mapped the complex problem of an electron moving through a crystal experiencing very complicated potentials to a textbook-type "particle in a well-defined potential" problem, which is solvable. The particle mass is renormalized, absorbing the details of the crystal potential. The *real* wavefunction of the wavepacket that models the particle-like nature of the electrons is given by $\psi(r) \approx u_{n0}(r)C(r)$, where $u_{n0}(r)$ is the periodic part of the Bloch eigenstates of the crystal that result from the periodic crystal potential. However, the beauty of the effective mass approximation is that the envelope function is all that is needed to find the bandstructure of the low-dimensional structures[2]! The envelope function concept is a powerful tool, as is demonstrated in its use in determining bandstructure modifications due to quantum confinement of carriers in low-dimensional structures.

14.3 3D bulk bandstructure

In a bulk semiconductor in the absence of external fields, $V(r) + E_c(r) = E_{c0}$ is a constant energy (flatband conditions), and thus the solution of

[1] An example is the semiconductor (diode) laser. The first semiconductor lasers were band-engineered by doping (i.e., they were p-n junctions). Heterostructures were then used to increase their efficiency. We discuss lasers in Chapter 29.

[2] Note that the bulk bandstructure is assumed to be known. The effective mass contains information about the bulk bandstructure.

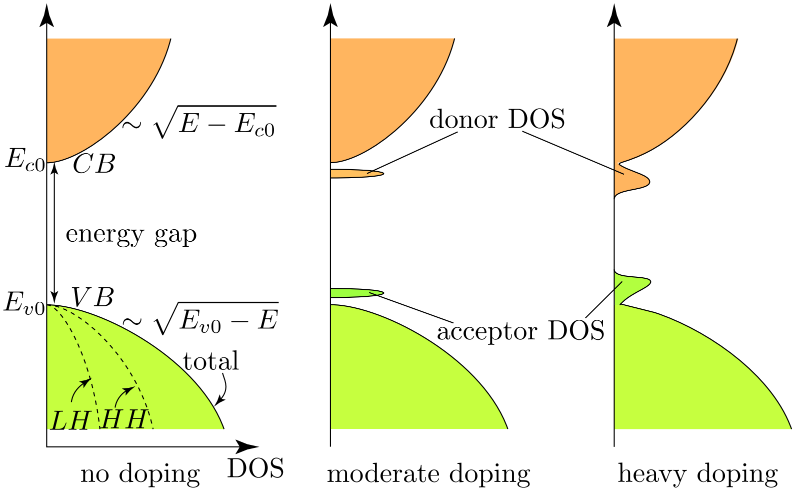


Fig. 14.4 Density of states of bulk (undoped), moderately doped and heavily doped semiconductors.

the effective mass equation yields envelope functions

$$C(r) = \frac{1}{\sqrt{V}} e^{i\vec{k}\cdot\vec{r}}, \tag{14.19}$$

and energies

$$E(k) = E_{c0}(r) + \frac{\hbar^2 k^2}{2m^\star} = E_{c0}(r) + \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_{xx}^\star} + \frac{k_y^2}{m_{yy}^\star} + \frac{k_z^2}{m_{zz}^\star}\right). \tag{14.20}$$

One should not forget that even though the k are written as continuous variables, they are actually quantized, taking values

$$k_x = k_y = k_z = \frac{2\pi}{L} m, \tag{14.21}$$

where $m = 0, \pm 1, \pm 2, \ldots$. Since L is a macroscopic length, the quantization is very fine, and for practical purposes, k can be assumed continuous.

The density of states (DOS) is given by (from Section 5.5),

$$g_{3D}(E) = \frac{1}{2\pi^2}\left(\frac{2m^\star}{\hbar^2}\right)^{3/2} \sqrt{E - E_{c0}}, \tag{14.22}$$

from which one gets a carrier concentration in the conduction band

$$n = \int_0^\infty dE f_{FD}(E) g_{3D}(E) = N_C^{3D} F_{1/2}\left(\frac{E_C - E_F}{k_B T}\right) \approx N_C^{3D} e^{-\frac{E_C - E_F}{k_B T}}, \tag{14.23}$$

where $F_j(\ldots)$ is the Fermi–Dirac integral function. The approximation holds only when the Fermi–Dirac distribution can be approximated by a Maxwell–Boltzmann form. Here, it is easily shown that N_C^{3D} is a effective band-edge DOS[3]:

[3] Note the similarity to the expressions derived in Chapter 5, with the substitution $m_e \to m^\star$, which is the essence of the effective mass method.

$$N_C^{3D} = 2(\frac{m^\star k_B T}{2\pi\hbar^2})^{\frac{3}{2}}. \tag{14.24}$$

Similar results hold for valence bands, where the contributions from the light and heavy hole bands add to give the total DOS. This is shown schematically in Fig. 14.4.

14.4 Doped semiconductors

The radius a_B^0 and ionization energy Ry^0 of an electron in the ground state of a hydrogen atom are

$$\boxed{a_B^0 = \frac{4\pi\epsilon_0\hbar^2}{m_e q^2}} \text{ and } \boxed{Ry^0 = \frac{1}{2}\frac{q^2}{4\pi\epsilon_0 a_B^0} = \frac{m_e q^4}{2(4\pi\epsilon_0)^2\hbar^2},} \tag{14.25}$$

where m_e is the free–electron mass and q is the electron charge. Doping a semiconductor adds states in the bandgap of the semiconductor. A shallow dopant adds states close to the band-edges. Considering a shallow donor, the hydrogenic-model solution from the effective mass equation (see Fig. 14.5 for an illustration)

$$[-\frac{\hbar^2}{2m^\star}\nabla^2 - \frac{q^2}{4\pi\epsilon r}]C(r) = (E - E_c)C(r) \tag{14.26}$$

shows that the eigenvalues are similar to that of a hydrogen atom, given by $E_n = E_{c0} - Ry^\star/n^2$, where $Ry^\star = Ry^0 \times (m^\star)/\epsilon_r^2$ is the modified hydrogenic energy levels, with Ry^0 the Rydberg energy, or the ground-state ionization energy. The ground-state envelope functions around the donor atoms given by

$$C(r) \sim e^{-r/r_0} \tag{14.27}$$

is spread over many lattice constants ($r_0 = a_B^0(\epsilon_r/m^\star) \gg a$). This implies that in k-space, the donor states are localized to $\Delta k \sim 1/r_0$. If the donor electron envelope function is spread over 1000 atoms in real space, in k-space it will be restricted to $\sim 1/1000$ of the volume of the Brillouin zone. Thus, for all practical purposes, the donor states are assumed to be "atomic-like". Energy separations between these individual atomic-like states is very small.

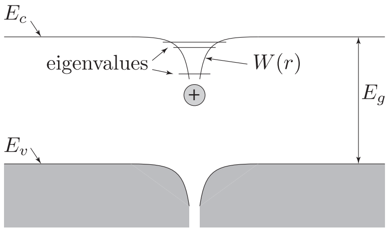


Fig. 14.5 The effective mass picture of an ionized donor in a semiconductor. The periodic potential of the semiconductor creates band edges E_c and E_v, in which electrons and holes of respective effective masses can move. An ionized donor creates an additional Coulomb potential $W(r)$ which creates additional bound state eigenvalues for electrons in the conduction band, similar to the potential of a proton experienced by a free–electron in a hydrogen atom. In Chapter 20, Section 20.5 we discuss how the eigenvalues of donor states can be used to make quantum-bits (qubits).

For heavy doping however, many changes can occur. The adjacent radii of electrons associated with adjacent donors can overlap, leading to formation of **impurity bands**. Then, the semiconductor acquires metal-like properties, since thermal activation of carriers into the bands is not necessary for transport. The effects of moderate and heavy doping on the DOS of bulk semiconductors is shown in Fig. 14.4.

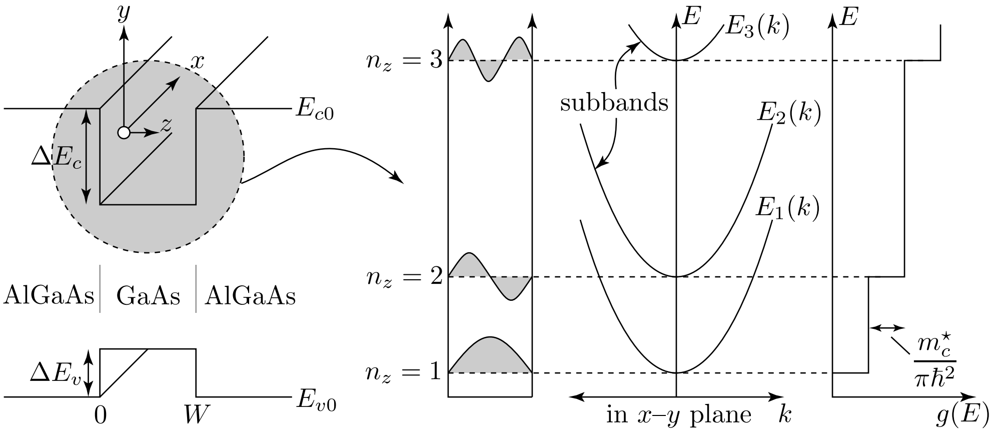


Fig. 14.6 Bandstructure, and DOS of realistic heterostructure quantum wells.

14.5 2D quantum wells

Quantum wells are formed by sandwiching a thin layer of semiconductor between barrier layers of wider bandgap. The finite extent of the quantum well layer makes the conduction band profile mimic a one-dimensional quantum well in the direction of growth (z-direction), leaving motion in the $x-y$ plane free (Fig. 14.6). Thus, the square-well potential (with reference to the conduction band edge E_{c0}) is written as

$$V(x,y,z) = 0, z < 0 \tag{14.28}$$

$$V(x,y,z) = 0, z > W \tag{14.29}$$

$$V(x,y,z) = -\Delta E_c, 0 \leq z \leq W. \tag{14.30}$$

Using the effective mass equation with this potential, it is evident that the envelope function should decompose as

$$C_{n_z}(x,y,z) = \phi(x,y)\chi_{n_z}(z) = [\frac{1}{\sqrt{A}}e^{i(k_x x + k_y y)}] \cdot [\chi_{n_z}(z)]. \tag{14.31}$$

If the quantum well is assumed to be infinitely deep, by simple wave-fitting procedure the z-component of the electron quasi-momentum is quantized to the values[4]

$$k_{n_z} = \frac{\pi}{W} n_z, \tag{14.32}$$

where $n_z = 1, 2, 3, \ldots$. From simple particle-in-a-box model in quantum mechanics, the normalized z-component of the envelope function is

$$\chi_{n_z}(z) = \sqrt{\frac{2}{W}} \sin \frac{\pi n_z z}{W}. \tag{14.33}$$

[4] Only waves that satisfy $n_z(\lambda/2) = W$ fit into the well of width W, leading to $k_{n_z} = 2\pi/\lambda = (\pi/W)n_z$.

The bandstructure is the set of energy eigenvalues obtained from the effective mass equation, given by

$$E(k) = E_{c0} + \underbrace{\frac{\hbar^2}{2}\left(\frac{k_x^2}{m_{xx}^\star} + \frac{k_y^2}{m_{yy}^\star}\right)}_{E_{2D}(k_x,k_y)} + \underbrace{\frac{\hbar^2}{2m_{zz}^\star}\left(\frac{\pi n_z}{W}\right)^2}_{E_{1D}(n_z)}, \tag{14.34}$$

which evidently decomposes to a free-electron component in the $x-y$ plane and a quantized component in the z-direction. The bandstructure consists of multiple bands $E_{2D}(k_x, k_y)$, each indexed by the quantum number n_z; this is shown in Fig. 14.6.

The DOS of electrons confined in an ideal 2D plane is a constant, given by $g_{2D}(E) = m^\star/\pi\hbar^2$ (from Section 5.4). In the quantum well, each subband corresponding to an n_z is an ideal 2-D system, and each subband contributes $g_{2D}(E)$ the total DOS. This is shown schematically in Fig. 14.6. Thus, the DOS of the quantum well is

$$g_{QW}(E) = \frac{m^\star}{\pi\hbar^2}\sum_{n_z}\theta(E - E_{n_z}), \tag{14.35}$$

where $\theta(\ldots)$ is the unit step function. The carrier density in the first subband of an ideal 2D electron system is thus given by

$$n_{2D} = \int_0^\infty dE f_{FD}(E) g_{2D}(E) = \underbrace{\frac{m^\star k_B T}{\pi\hbar^2}}_{N_C^{2D}} \ln(1 + e^{\frac{E_F - E_1}{k_B T}}), \tag{14.36}$$

where E_1 is the ground state energy, E_F is the Fermi level, and N_C^{2D} is the effective band-edge DOS, the 2-dimensional counterpart of N_C^{3D} defined in Equation 14.24. Several examples were given earlier in Chapter 5 and elsewhere.

For the quantum well, which houses many subbands, the DOS becomes a sum of each subband (Fig. 14.6), and the total carrier density is thus a sum of 2D-carriers housed in each subband –

$$n_{2D} = \sum_j n_j = N_c^{2D}\sum_j \ln(1 + e^{\frac{E_F - E_j}{k_B T}}). \tag{14.37}$$

Note that for a 2D system, no approximation of the Fermi–Dirac integral is necessary, for the carrier density is obtained analytically.

It is important to note that the confining potential in the z-direction can be engineered almost at will by modern epitaxial techniques by controlling the spatial changes in material composition. For example, a popular quantum well structure has a parabolic potential ($V(z) \sim z^2$), which leads to the E_{n_z} values spaced in equal energy intervals – this is a characteristic of a square, or harmonic oscillator potential. Another extremely important quantum well structure is the triangular well potential ($V(z) \sim z$), which appears in MOSFETs, HEMTs, and

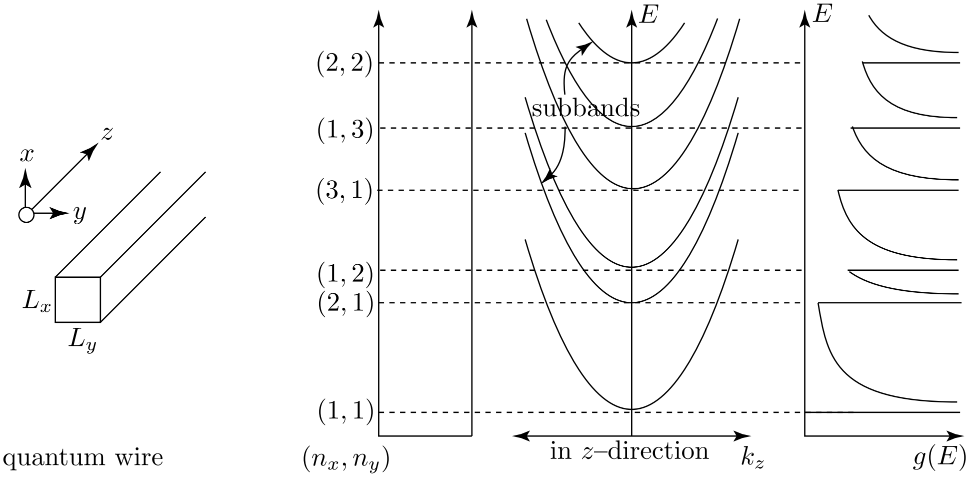


Fig. 14.7 Bandstructure, and DOS of realistic quantum wires.

quantum wells under electric fields. The triangular well leads to E_{n_z} values given by Airy functions. Regardless of these details specific to the *shape* of the potential, the bandstructure and the DOS remain similar to the square well case; the only modification being the E_{n_z} values, and the corresponding subband separations.

14.6 1D quantum wires

Artificial quantum wires are formed either lithographically (top-down approach), or by direct growth in the form of semiconductor nanowires or nanotubes (bottom-up approach). In a quantum well, out of the three degrees of freedom for real space motion, carriers were confined in one direction, and were free to move in the other two directions. In a quantum wire, electrons are free to move freely in one direction only (hence the name "wire"), and the other two degrees of freedom are quantum-confined. Assume that the length of the wire (total length L_z) is along the z-direction (see Fig. 14.7), and the wire is quantum-confined in the $x - y$ plane ($L_x, L_y \ll L_z$). Then, the envelope function naturally decomposes into

$$C(x, y, z) = \chi_{n_x}(x) \cdot \chi_{n_y}(y) \cdot (\frac{1}{\sqrt{L_z}} e^{ik_x x}), \tag{14.38}$$

and the energy eigenvalues are given by

$$E(n_x, n_y, k_z) = E(n_x, n_y) + \frac{\hbar^2 k_k^2}{2m_{zz}^\star}. \tag{14.39}$$

If the confinement in the $x - y$ directions is by infinite potentials (a useful approximation applicable in many quantum wires), then similar to the quantum well situation, a wave-fitting procedure gives

$$k_{n_x} = \frac{\pi}{L_x} n_x \text{ and } k_{n_y} = \frac{\pi}{L_y} n_y, \tag{14.40}$$

where $n_x, n_y = 1, 2, 3, \ldots$ independently.

The eigenfunctions assume the form

$$C_{n_x,n_y}(x,y,z) = [\sqrt{\frac{2}{L_x}} \sin(\frac{\pi n_x}{L_x} x)] \cdot [\sqrt{\frac{2}{L_y}} \sin(\frac{\pi n_y}{L_y} y)] \cdot [\frac{1}{\sqrt{L_z}} e^{ik_x x}], \tag{14.41}$$

and the corresponding bandstructure is given by

$$E(n_x, n_y, k_z) = \underbrace{[\frac{\hbar^2}{2m_{xx}}(\frac{\pi n_x}{L_x})^2] + [\frac{\hbar^2}{2m_{yy}}(\frac{\pi n_y}{L_y})^2]}_{E(n_x,n_y)} + \frac{\hbar^2 k_z^2}{2m_{zz}^\star}. \tag{14.42}$$

Multiple subbands are formed, similar to the quantum well structure. A new subband forms at each eigenvalue $E(n_x, n_y)$, and each subband has a dispersion $E(k_z) = \hbar^2 k_z^2 / 2m_{zz}$ (Fig. 14.7).

The DOS of electrons confined to an ideal 1D potential is given by (from Section 5.3),

$$g_{1D}(E) = \frac{1}{\pi}\sqrt{\frac{2m^\star}{\hbar^2}} \frac{1}{\sqrt{E - E_1}}, \tag{14.43}$$

where E_1 is the lowest allowed energy (ground state). Due to multiple subbands, the DOS acquires peaks at every eigenvalue $E(n_x, n_y)$. Since there are two quantum numbers involved, some eigenvalues can be degenerate, and the peaks can occur at irregular intervals as opposed to the quantum well case. The general DOS for a quantum wire can thus be written as

$$g_{QWire}(E) = \frac{1}{\pi}\sqrt{\frac{2m^\star}{\hbar^2}} \sum_{n_x,n_y} \frac{1}{\sqrt{E - E(n_x, n_y)}}, \tag{14.44}$$

which is shown schematically in Fig. 14.7.

14.7 0D quantum dots

The quantum dot is the ultimate nanostructure. All three degrees of freedom are quantum confined; therefore there is no plane-wave component of electron wavefunctions. The envelope function for a "quantum box" of sides L_x, L_y, L_z (see Fig. 14.8) is thus written as

$$C(x,y,z) = \chi_{n_x}(x)\chi_{n_y}(y)\chi_{n_z}(z), \tag{14.45}$$

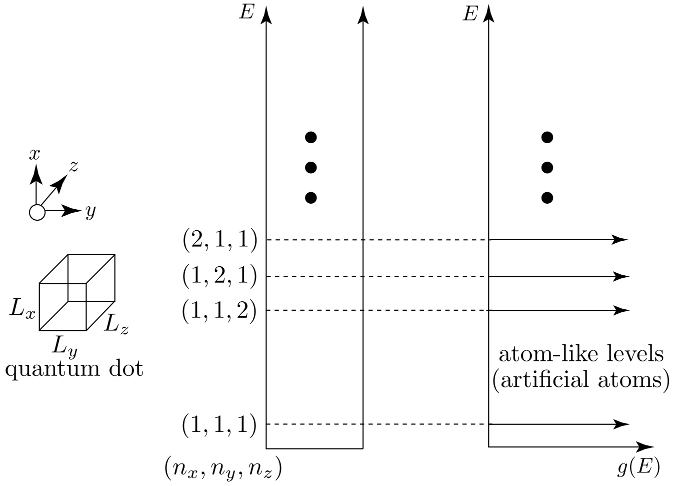


Fig. 14.8 Energy levels and DOS of quantum dots.

and if the confining potential is infinitely high, we have $k_{n_i} = (\pi/L_i)n_i$ for $i = x, y, z$. The envelope functions are thus given by

$$C(x,y,z) = [\sqrt{\frac{2}{L_x}}\sin(\frac{\pi n_x}{L_x})] \cdot [\sqrt{\frac{2}{L_y}}\sin(\frac{\pi n_y}{L_y})] \cdot [\sqrt{\frac{2}{L_z}}\sin(\frac{\pi n_z}{L_z})], \tag{14.46}$$

and the energy eigenvalues are given by

$$E(n_x, n_y, n_z) = \frac{\hbar^2}{2m_{xx}}(\frac{\pi n_x}{L_x})^2 + \frac{\hbar^2}{2m_{yy}}(\frac{\pi n_y}{L_y})^2 + \frac{\hbar^2}{2m_{zz}}(\frac{\pi n_z}{L_z})^2. \tag{14.47}$$

Note that the energy eigenvalues are no more quasi-continuous, and are indexed by three quantum numbers (n_x, n_y, n_z). Thus, it does not make sense to talk about "bandstructure" of quantum dots; the DOS is a sum of delta functions, written as

$$g_{QDot} = \sum_{n_x, n_y, n_z} \delta(E - E_{n_x, n_y, n_z}). \tag{14.48}$$

This is shown schematically in Fig. 14.8. Since there is no direction of free motion, there is no transport *within* a quantum dot, and there is no quasi-continuous momentum components[5]. Fabricating quantum dots by lithographic techniques is pushing the limits of top-down approach to the problem. On the other hand, epitaxial techniques can coax quantum dots to self-assemble by cleverly exploiting the strain in lattice-mismatched semiconductors. On the other hand, bottom-up techniques of growing nanocrystals in solution by chemical synthetic routes is becoming increasingly popular.

[5] The sharp energy levels of atoms in the periodic table are fixed, but those in quantum dots are up to us to design: they are **artificial atoms**!

14.8 Finite barrier heights

Fig. 14.9 shows how the eigenvalues change for a finite barrier height for quantum wells, in direct comparison to the eigenvalues for the ∞-barrier height case. The solutions are obtained directly by enforcing the continuity of the wavefunction and its derivative at the two walls of the barrier. The procedure is discussed in Exercise 14.2. Panel (a) shows a schematic depiction of the finite quantum well, with the expected bound state eigenvalues and the corresponding eigenfunctions. Figs. 14.9 (b) and (c) show the eigenvalues with the solid lines as the finite barrier height well eigenvalues $E_1, E_2, ...$, and the dashed lines the ∞-well values $E_1^{\infty}, E_2^{\infty}, ...$ The striking features are (a) the drop in energy eigenvalues with increasing quantum well thickness, (b) the energy eigenvalues for a finite well are smaller than the ∞-barrier well approximation, (c) the eigenvalues for finite well heights gradually approach that of ∞-well, and (d) that the finite wells have a **finite number of bound states**, unlike the ∞-well. The lower energy eigenvalues are expected because of the penetration of the electron wavefunction into the barrier by tunneling. Because of this, the effective width of the finite well is similar to a thicker ∞-well. The finite number of bound states are also expected because any electron energy larger than the barrier height is a free state that is a part of the continuum with plane-wave solutions. For example, for $L_w = 15$ nm with $U_0 = 0.3$ eV barrier height and $m^{\star} = 0.067 m_e$ representative of the conduction band effective mass of GaAs, there are four bound states in the quantum well.

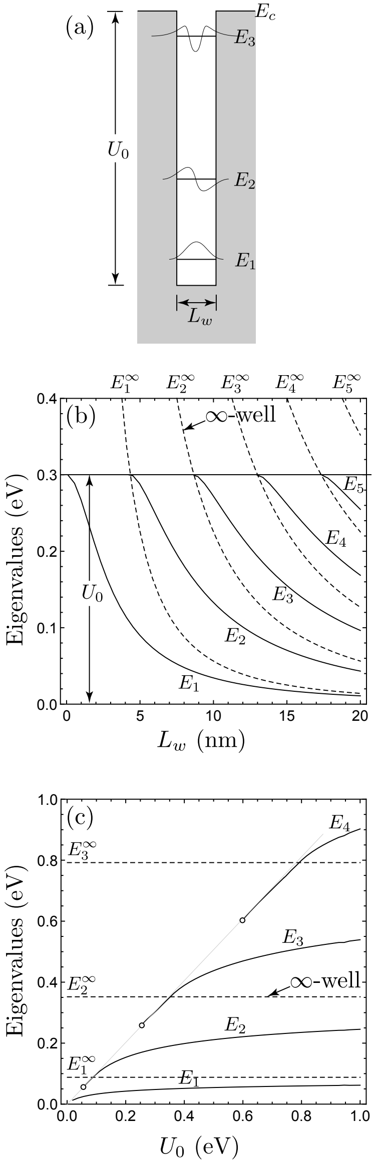


Fig. 14.9 Eigenvalues of a quantum well with finite barrier height. $m^{\star} = 0.067 m_e$ is assumed, and (c) is for a fixed L_w.

14.9 Multilayers and superlattices

Consider two quantum wells located close to each other in space. If the potential barrier separating them has an ∞ height but finite thickness, the states in the two wells are completely isolated from each other because the wavefunction must go to zero when the potential goes to ∞. This is necessary for the terms in the Schrödinger equation to remain finite. But a finite barrier height allows for the wavefunction from one well to "leak" into the other. This property is heavily used in semiconductor quantum structures called **multilayers** and **superlattices**.

Fig. 14.9 (a) indicated the penetration of the envelope wavefunction into the finite barriers. Consider now[6] in Fig. 14.10 (a) the numerically calculated conduction band edge E_c for an $Al_{0.2}Ga_{0.8}As$ / GaAs / $Al_{0.2}Ga_{0.8}As$ heterostructure with $q\phi_B = 0.6$ eV barrier heights at the left and right ends. The GaAs quantum well is $L_w = 4$ nm thick and is located at the center. The $Al_{0.2}Ga_{0.8}As$ layers are doped for 50 nm at $N_d = 10^{17}/\text{cm}^3$ around the quantum well. For Figs. 14.10 (b – f) the same $Al_{0.2}Ga_{0.8}As$ layer scheme is used, but the number of quantum wells is increased from $N = 1$ to $N = 10$. The barrier layers separating the quantum wells are $Al_{0.2}Ga_{0.8}As$. Both the quantum wells and the barriers separating them are undoped.

---

[6] The following numerical calculations in this section are performed using a self-consistent solution of the Schrödinger equation and the Poisson equation (see Section 15.6).

Fig. 14.10 (a) shows the energy eigenvalue E_1 for the lowest energy bound state in the quantum well in the z-direction, and the corresponding envelope function $C_1(z)$ where the penetration to the barriers is seen. In such heterostructures typically the $x - y$ direction is uniform and the electron envelope function is a 2D plane wave in the $x - y$ plane, but this point is not the subject of the following discussion. Here we focus primarily on the potential for motion along the z-direction by coupling between the adjacent wells.

Fig. 14.10 (b) shows the situation for $N = 2$ quantum wells. The lowest eigenvalues now split into two: E_1 and E_2, and the corresponding envelope functions $C_1(z)$ takes a ″bonding′, and $C_2(z)$ an ″antibonding″ character. Note that $C_2(z)$ has one zero-crossing. Also note that for $N = 2$ wells the envelope function $C_1(z)$ for the lower energy state looks like the **sum** of the envelope functions for the $N = 1$ case but centered at each well. Similarly $C_2(z)$ for $N = 2$ wells looks like the **difference** of the envelope functions of the $N = 1$ case centered at each well. Notice the analogy to Fig. 11.3 for the corresponding states in atoms and molecules. The $N = 2$ coupled quantum well exhibits a similar feature as individual atoms, except that here it is for a quantum well in a semiconductor crystal: the envelope functions $C_1(z)$ and $C_2(z)$ have to be multiplied by the periodic Bloch functions. The lateral $x - y$ plane wave motion is another major difference from the atomic situation.

Figs. 14.10 (c – f) show successively larger number of coupled quantum wells. The envelope function $C_i(z)$ corresponding to the E_i^{th} eigenvalue has $i - 1$ nodes. The reader is encouraged to examine the envelope functions and draw analogies to the tight-binding models that were described in Chapter 11, Fig. 11.6, while keeping in mind that here we are discussing envelope functions. For the $N = 10$ case shown in Fig. 14.10 (f), the $C_{10}(z)$ envelope function has 9 nodes. The lower energy states for $N = 10$ are sufficiently delocalized in the z-direction. This implies they can move relatively freely between the quantum wells, but are still bound by the larger barriers outside of the wells. There are additional energy eigenvalues that have energies larger than the barrier height of the quantum wells. These states are bound by the wider barriers formed between the cladding $Al_{0.2}Ga_{0.8}As$ layers. Since the electron in these states can move over a larger z-distance than those inside the quantum wells, their energies will be spaced closely.

Furthermore, it will be noted that as as the number of coupled quantum wells N increases, more energy states are clumped into a finite energy width. Since this fact is not that clear in Fig. 14.10, the N lowest energy eigenvalues calculated for N coupled quantum wells are plotted separately in Fig. 14.11 for $N = 1$ to $N = 16$ wells. It is now seen explicitly that, for example, the maximum minus minimum energy eigenvalues approaches a **finite energy width** as N increases.

Imagine now that the number of coupled quantum wells increases to a large number $N \to \infty$. In that case, the energy eigenvalues become

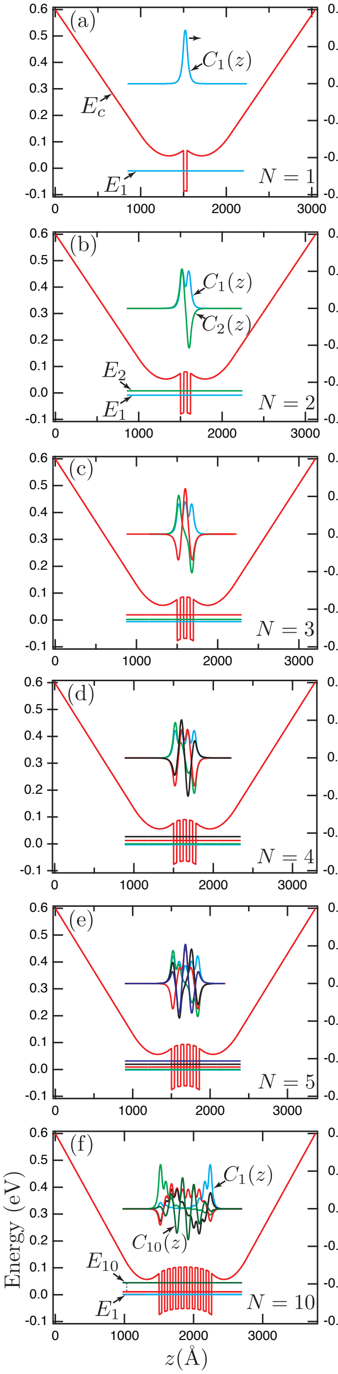


Fig. 14.10 Energy band diagrams of the conduction band of AlGaAs/GaAs heterostructures. The number of quantum wells range from $N = 1$ to $N = 10$. The energy eigenvalues E_i and corresponding eigenfunctions $\psi_i(z)$ are shown.

very closely spaced. Furthermore, similar to the tight binding model of Fig. 11.6 for the s-band, there is a energy band formed for motion in the z-direction by the combination of the ground states of the wells. This is called the **miniband**, and this crystal is called a **superlattice**. Notice that this superlattice has a lattice constant along the z-direction that is a multiple of the lattice constant of the semiconductor crystal itself. Therefore, as opposed to the lattice constant of the semiconductor crystal (e.g. GaAs or $Al_{0.2}Ga_{0.8}As$ here), the effective lattice constant of the miniband is $L_w + L_b$, the sum of the well and barrier thicknesses. Thus, the Brillouin-zone edges of the artificial superlattice for motion along the z-direction is therefore $-\pi/(L_w + L_b) \leq k_z \leq +\pi/(L_w + L_b)$.

There can be a superlattice miniband formed for each energy eigenvalue of the quantum well. For example, if the single well allowed a second bound state, this state could couple more strongly with nearby wells because of a smaller barrier height. Therefore, the second miniband with larger bandwidth and smaller effective mass can be formed, as is indicated schematically in Fig. 14.12.

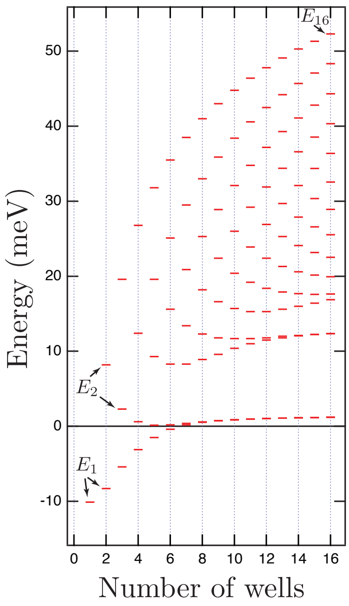


Fig. 14.11 Lowest energy eigenvalues as a function of the number of coupled wells.

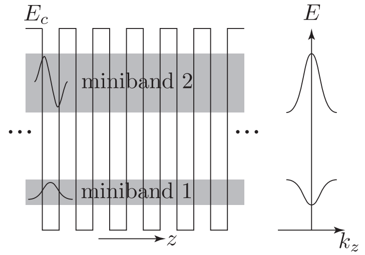


Fig. 14.12 Minibands formed in a semiconductor superlattice, and the corresponding energy dispersion in the z direction. Such superlattices are typically used in semiconductor electronic or photonic devices for carrier transport

14.10 Wannier functions

Because Bloch eigenstates $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r})$ are extended over the entire crystal, to handle carriers that are confined to fixed regions in space, in this chapter we have developed linear combinations of the Bloch $\mathbf{k}$-states to create the effective mass wavefunction $\psi(\mathbf{r}) \approx C(\mathbf{r})u_{\mathbf{k}}(\mathbf{r})$. Since the envelope function $C(\mathbf{r})$ is a slowly varying function, we realized that it is really the solution to the band-edge profile, and it can handle quantum confinement in a simple, transparent, and intuitive way. In fact that is the reason the effective mass theory is so popular as a design tool.

Instead of extended Bloch functions, to which we arrived at by starting from free–electrons, one can ask if there is a *complete* basis set of mutually orthogonal wavefunctions that are *localized* on atomic sites. The idea here is similar in spirit to the tight-binding approach to bandstructure of approaching electronic states from the molecular orbitals tied to each atom, rather than the nearly free–electron model, or the pseudopotential models. Such a localized, orthogonal, and complete set of functions was introduced by Wannier. The Wannier function $w_n(r - R)$ of the n^{th} band tied to the lattice site R is related to the Bloch function in the following way:

$$w_n(r - R) = \frac{1}{\sqrt{N}} \sum_{\mathbf{k}\in\text{BZ}} e^{-i\mathbf{k}\cdot\mathbf{R}}\psi_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{k}\in\text{BZ}} e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})}u_{n\mathbf{k}}(\mathbf{r}), \tag{14.49}$$

which is a summation over $\mathbf{k}$-states in the entire BZ. The periodic part of the Bloch function is related to the Wannier function via

$$u_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{N}} \sum_{\mathbf{R}} e^{-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{R})} w_n(r - R). \tag{14.50}$$

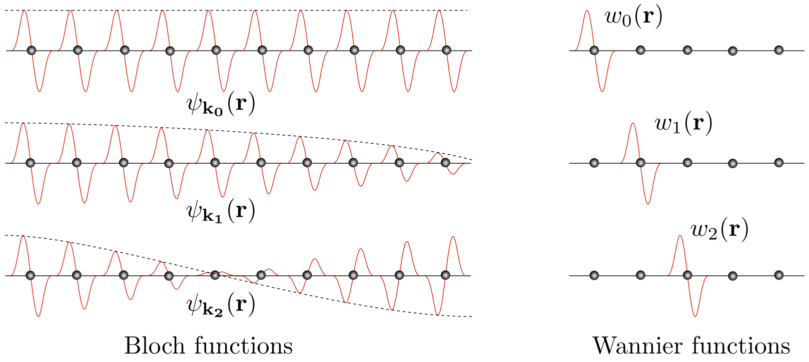

$\psi_{\mathbf{k_0}}(\mathbf{r})$
$\psi_{\mathbf{k_1}}(\mathbf{r})$
$\psi_{\mathbf{k_2}}(\mathbf{r})$
Bloch functions
$w_0(\mathbf{r})$
$w_1(\mathbf{r})$
$w_2(\mathbf{r})$
Wannier functions

Fig. 14.13 Real parts of Wannier functions compared to Bloch functions.

Fig. 14.13 highlights the difference between Bloch and Wannier functions. While the Bloch functions are extended, the Wannier functions are localized at atomic sites. Though Bloch states are energy eigenvalues for the electron in the crystal, the Wannier states are not. The reason we are introducing the Wannier states here is to highlight and contrast them with both the Bloch states, and also the effective mass states[7]. Effective mass states are linear combinations of Bloch states, and are also linear combinations of Wannier states. Since the bandstructure $E(\mathbf{k})$ is known from experiments, the Bloch states are used as the basis for effective mass theory. However, for certain physical phenomena, the Bloch states are not well suited. An example of such phenomena is the presence of spontaneous and piezoelectric polarization in semiconductors. The reason is that such phenomena classically are thought of as charge dipoles localized in every unit cell, and combinations of Bloch functions cannot easily provide something that is localized. The Wannier function is superior to explain such phenomena.

Fig. 14.14 Gregory Wannier introduced the concept of spatially orthogonal wavefunctions in crystals.

14.11 Chapter summary section

In this chapter, we learned:

- How the effective mass theory maps the complex problem of electrons in periodic or quantum confined potentials into the simplest problems of quantum mechanics, such as the hydrogen atom or the various particle in a box problems, and
- How to exploit this connection to understand, and design semiconductor doping, and heterostructure based quantum wells, wires, and dots.

---

[7] An additional significance of Wannier functions is that they can be defined separately for each band, whereas wavepackets from Bloch functions require basis elements from multiple bands.

Further reading

The Physics of Low Dimensional Semiconductors by Davies provides an excellent treatment of the effective mass method, and its applications to several quantized heterostructures with beautiful illustrations and examples. **Wave Mechanics Applied to Semiconductor Heterostructures** by Bastard is a classic text that explains several intricacies of the quantum mechanics of heterointerfaces using effective mass methods.

Exercises

(14.1) **Effective mass methods: doping and quantum dots**

We derived the effective mass approximation in this chapter, in which the complicated problem of a *free–electron* in a periodic potential was mapped to a much simpler problem of an electron with an *effective mass* in a free potential. The key step was to create a wavepacket by constructing a linear combination of Bloch eigenstates $\psi_k(r) = u_k(r)e^{ikr}$ in the form $\phi_k(r) = \int \frac{dk}{2\pi} C(k) u_k(r) e^{ikr}$, and using Fourier transform properties with the Hamiltonian operator. The effective mass equation in the presence of *perturbations* to the periodic crystal potential then is written as

$$[E_n(-i\nabla) + W]C(r) = EC(r), \tag{14.51}$$

where $C(r)$ is the *envelope* function, and $E_n(-i\nabla)$ is an *operator* obtained from the bandstructure $E_n(k)$ of band n by replacing $k \to -i\nabla$. W is a time-dependent, or time-independent *perturbation* to the periodic potential of the crystal. Instead of the Bloch functions $\psi_k(r)$, we can now work with the envelope functions $C(r)$, remembering that the wavefunction of the wavepacket is the product of the envelope function and the periodic part of the Bloch-function, i.e., $\phi(r) = C(r)u_k(r)$. For this problem, consider the conduction band with a parabolic bandstructure characterized by an effective mass $m_c^\star$ and band-edge E_c, such that $E_c(k) = E_c + \frac{\hbar^2 k^2}{2m_c^\star}$.

(a) Show that the effective mass equation is then

$$[-\frac{\hbar^2}{2m_c^\star}\nabla^2 + W]C(r) = (E - E_c)C(r). \tag{14.52}$$

Note that this is in the form of a modified Schrödinger equation, and is referred to as the effective mass Hamiltonian. Show that the solutions in the absence of the perturbation are simple plane waves, $C_k(r) = \frac{1}{\sqrt{V}}e^{ikr}$. Find k.

(b) **Doping**: When we introduce a dopant atom in the semiconductor (see Fig. 14.16), the perturbation due to a single donor atom in a 3-Dimensional crystal semiconductor is a Coulomb potential $W(r) = -\frac{e^2}{4\pi\epsilon_s r}$ with ϵ_s the dielectric constant of the semiconductor. Argue that this effective mass problem maps exactly to the hydrogen atom problem, and show that the energy of the shallow donors are $E_n = E_c - Ry^\star/n^2$, where $Ry^\star = \frac{(m_c^\star/m_e)}{(\epsilon_s/\epsilon_0)^2} Ry^0$, where $Ry^0 = -13.6$ eV is the

Fig. 14.15 Walter Kohn with Joaquin Luttinger developed the effective mass theory for semiconductors. The effective mass theory arms us with a powerful method to design semiconductor nanostructures and enormously simplifies the design of quantum wells, wires, and dots. Kohn was awarded the Nobel Prize in 1998 for the development of density functional theory (DFT).

ground state energy of an electron in the hydrogen atom. Also show that the radius of the donor electron is modified from the hydrogen electron radius to $a_B^\star = \frac{\epsilon_s/\epsilon_0}{m_c^\star/m_e} a_B^0$, where $a_B^0 = 0.53$ Angstrom is the Bohr radius of the electron in the hydrogen atom. Estimate the ionization energy of the donor, and the radius in a semiconductor with $\epsilon_s = 10\epsilon_0$ and $m_c^\star = 0.1 m_e$. From these considerations, argue why bands with heavy effective masses may be difficult to dope.

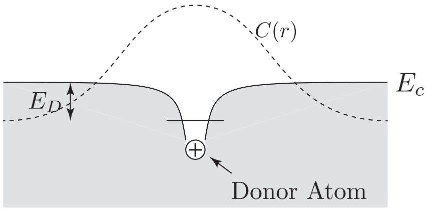


Fig. 14.16 Schematic representation of a charged donor, its activation energy E_D and the envelope function of the donor electron $C(r)$.

(c) **Quantum Dots**: Suppose we have a narrow bandgap semiconductor quantum dot of size $L_x = L_y = L_z = L$ embedded in a wide-bandgap semiconductor matrix. Assume the conduction band offset ΔE_c and the valence band offset ΔE_v are very large, such that an electron in the conduction band and holes in the valence band of the quantum dot effectively see infinitely tall barriers. Find the allowed energies of the electrons and hole states in the quantum dot as a function of the dot size L, and the conduction and valence band effective masses $m_c^\star$ and $m_v^\star$. If the bulk bandgap of the narrow bandgap semiconductor is E_g, what is the energy of a photon that will be emitted if an electron transitions from the CB ground state to the VB ground state? Make a plot of the emitted photon energy as a function of the quantum dot size from 1 nm $\leq L \leq$ 10 nm, for the following parameters of the narrow bandgap semiconductor: $m_c^\star = m_v^\star = 0.1 m_e$, $E_g = 0.8$ eV. Fig. 14.17 shows how such quantum confinement is used to create photons of large energy for deep ultraviolet LEDs and lasers.

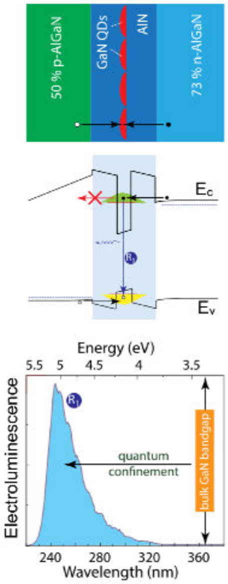


Fig. 14.17 A quantum dot UV light-emitting diode realized by blue-shifting the emission from GaN by quantum confinement.

(14.2) Quantum well heterostructures

The finite quantum well problem is the basis of all quantized structures based on compound semiconductor heterostuctures. In this problem you evaluate some examples to gain insight, and collect some *very* useful formulae for the quantum design of heterostructure devices.

(a) With relevant formulae and sketches, outline a graphical method for identifying the bound state eigenvalues and eigenfunctions in a finite quantum well of height U_0 and width L_w for a quantum well semiconductor material with effective mass $m^\star$. Show that the solution for allowed k values take the form $\sqrt{\frac{\theta_0^2}{\theta^2} - 1} = \tan\theta$ and $\sqrt{\frac{\theta_0^2}{\theta^2} - 1} = -\cot\theta$, where $\theta = \frac{kL_w}{2}$, and the characteristic constant $\theta_0^2 = \frac{m^\star L_w^2 U_0}{2\hbar^2}$.

(b) Show that in the case of a vanishingly small barrier height $U_0 \to 0$, there is still at least one bound state for the 1D quantum well with a *binding energy* equal to $\boxed{U_0 - E_1 \approx \theta_0^2 U_0.}$

(c) Show that the number of bound states is

$\boxed{N = 1 + \text{Int}[\frac{2\theta_0}{\pi}],}$ where $\text{Int}[x]$ is the largest integer smaller than x. Show that the numerical value is $N = 1 + \text{Int}[1.63(\frac{L_w}{1\text{ nm}})\sqrt{(\frac{m^\star}{m_0}) \cdot (\frac{U_0}{1\text{ eV}})}]$.

(d) Now consider the electron states in the conduction band of a heterostructure quantum well shown in Fig. 14.6, with $U_0 = \Delta E_c = 0.3$ eV, and $m_c^\star = 0.067 m_e$. How many bound states does a well of thickness $L_w = 5$ nm hold? Write the effective mass wavefunctions $C(r)$ for electrons in the ground state of the quantum well, and find its characteristic penetration depth into the barrier layer.

(e) Find the Fermi level in the quantum well if we fill the quantum well with electrons of 2D density $n_s = 10^{12}/\text{cm}^2$.

(14.3) A heterostructure step barrier

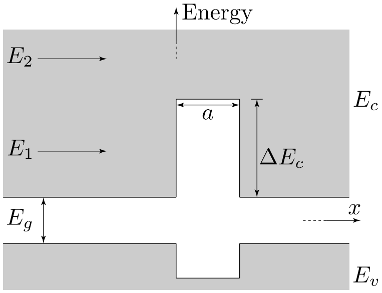


Fig. 14.18 A heterostructure step barrier.

Fig. 14.18 shows a step barrier for electrons formed at a heterojunction between two semiconductors of different bandgaps.

(a) Consider the 1D situation and the conduction band electron effective mass $m_c^\star$ to be identical in all regions. Show that the transmission probability for an electron incident from the left with kinetic energy $E = \frac{\hbar^2 k_l^2}{2m_c^\star}$ is

$$\boxed{T(E) = \frac{1}{1 + \frac{(\Delta E_c)^2}{4E(E - \Delta E_c)} \cdot \sin^2(k_b a)},} \tag{14.53}$$

where $k_b = \sqrt{\frac{2m_c^\star}{\hbar^2}(E - \Delta E_c)}$.

Solution: Let the electron wavevectors in the left, barrier, and right regions be k_l, k_b and k_r respectively. The wavefunctions are respectively $\psi_l(x) = e^{ik_l x} + r_l e^{-ik_l x}$, $\psi_b(x) = t_b e^{ik_b x} + r_b e^{-ik_b x}$ and $\psi_r(x) = t_r e^{ik_r x}$. Equating the wavefunctions and their 1st derivatives at $x = 0$, $\psi_l(0) = \psi_b(0)$, $\psi_l^{'}(0) = \psi_b^{'}(0)$ and at $x = a$, $\psi_l(a) = \psi_b(a)$, $\psi_l^{'}(a) = \psi_b^{'}(a)$ gives 4 equations in the 4 unknowns r_l, t_b, r_b, t_r. Note that each of these coefficients are in general complex quantities. Solving the 4 equations gives $r_l = \frac{r_{bl} - r_{br} e^{2ik_b a}}{r_{bl} r_{br} e^{2ik_b a} - 1}$, $t_r = \frac{t_{bl} e^{i(k_b - k_r)}}{r_{bl} r_{br} e^{2ik_b a} - 1}$, $t_b = \frac{t_l}{1 - r_{bl} r_{br} e^{2ik_b a}}$, and $r_b = \frac{t_l}{r_{bl} - \frac{1}{r_{br}} e^{-2ik_b a}}$. Here $r_{bl} = \frac{k_b - k_l}{k_b + k_l}$, $r_{br} = \frac{k_b - k_r}{k_b + k_r}$, $t_l = \frac{2k_l}{k_b + k_l}$, and $t_{bl} = \frac{4k_b k_l}{(k_b + k_l)(k_b + k_r)}$. From the amplitudes, the transmission probability across the barrier is given by $T(E) = |t_r|^2 = \frac{1}{1+x}$, the same as Equation 14.53.

(b) Find the reflection probability $R(E)$ as a function of the kinetic energy of the electron, and find an approximation for small barrier heights.

Solution: Since an electron incident from the left is either reflected or transmitted across the barrier layer to the right, the reflection probability and transmission probability must add to unity:

$$\boxed{R(E) = 1 - T(E) = \frac{1}{1 + \frac{4E(E - \Delta E_c)}{(\Delta E_c)^2} \cdot \csc^2(k_b a)}.} \tag{14.54}$$

This problem will be taken up in Chapter 24 since it is central to tunneling through barriers.

Carrier Statistics and Energy Band Diagrams

15

15.1 **Carrier statistics** 311
15.2 **E_F is constant at thermal equilibrium** 320
15.3 **Metal-semiconductor Schottky junctions** 320
15.4 **p-n homojunctions** 323
15.5 **Heterojunctions** 325
15.6 **Energy band diagrams: Poisson+Schrödinger** 327
15.7 **Polarization-induced doping in heterostructures** 330
15.8 **Chapter summary section** 331
Further reading 332
Exercises 333

Once a particular semiconductor structure is realized, the electron *energy bandstructures* $E(\mathbf{k})$ are locked in. As opposed to $E(\mathbf{k})$ which is the solution of the quantum mechanical energy eigenvalues of electrons in the $\mathbf{k}$-space, *energy band diagrams* tell us how the allowed electron energies vary in *real* space. The atomic composition of the semiconductor heterostructures, and the chemical doping determine how the mobile electron and hole concentrations, and the energy bands change in real space. In other words, they determine how the energy band edges around the bandgap $E_c(\mathbf{r})$ and $E_v(\mathbf{r})$ vary in real space. In this chapter, we develop the following core concepts essential towards controlling and manipulating electrons and holes in semiconductor electronic and photonic devices:

- What are the mobile carrier concentrations in an intrinsic, pure semiconductor?
- How can we vary the mobile carrier concentrations in semiconductors by factors of $\sim 10^{10}$ by doping?
- What are the carrier statistics and energy band diagrams in **homojunctions**: when the semiconductor is the same, but the doping varies in space?
- What are the carrier statistics and energy band diagrams in **heterojunctions**: when both the semiconductors *and* the doping vary in space?

15.1 Carrier statistics

Consider first a large piece of an intrinsic semiconductor. Assume it has no donor or acceptor doping. How many mobile electrons and holes can I measure in it at room temperature? We have already learned the carrier statistics relations: that the electron and hole densities in 3D are

$$n = N_c F_{\frac{1}{2}}\left(\frac{E_F - E_c}{k_b T}\right) \text{ and } p = N_v F_{\frac{1}{2}}\left(\frac{E_v - E_F}{k_b T}\right), \tag{15.1}$$

where $N_c = N_c^{3d} = g_s g_v^c (\frac{2\pi m_c^\star k_b T}{h^2})^{\frac{3}{2}}$ and $N_v = N_v^{3d} = g_s g_v^v (\frac{2\pi m_v^\star k_b T}{h^2})^{\frac{3}{2}}$. Note that E_F is the Fermi level, E_c is the conduction band edge, and

E_v is the valence band edge. The bandgap is $E_g = E_c - E_v$. The spin degeneracy is $g_s = 2$. The valley degeneracy will depend on the bandstructure $E(\mathbf{k})$ around the bandgap. For the simplest case, assuming there is only one conduction band minimum and one valence band maximum, $g_v^c = g_v^v = 1$, the band edge DOS are $N_c = 2(\frac{2\pi m_c^\star k_b T}{h^2})^{\frac{3}{2}}$ and $N_v = 2(\frac{2\pi m_v^\star k_b T}{h^2})^{\frac{3}{2}}$. Then, at equilibrium, since all the electrons in the conduction band in an intrinsic semiconductor must have come from thermal excitation of the valence band electrons across the bandgap, their numbers must be equal: $n = p = n_i$. The density n_i is called the **intrinsic carrier concentration**, and is a characteristic property of any semiconductor (Fig. 15.1). This intrinsic density is then given by

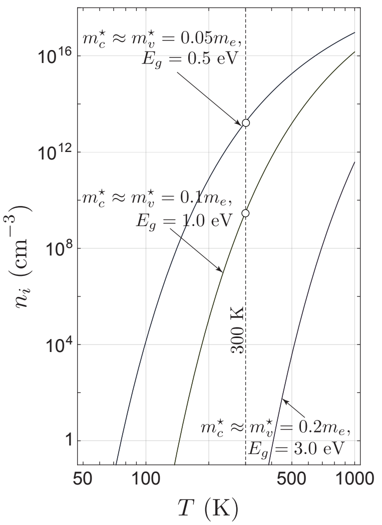


Fig. 15.1 Intrinsic carrier concentrations n_i vs. temperature for semiconductors with different effective masses and bandgaps. Note that at 500° C $\sim 773K$, intrinsic wide bandgap semiconductors (e.g. GaN or SiC) of an energy gap 3 eV have very few thermally generated electrons and holes compared to silicon, making wide bandgap semiconductors suitable for high-temperature electronics.

$$n = N_c F_{\frac{1}{2}}(\frac{E_F - E_c}{k_b T}) = N_v F_{\frac{1}{2}}(\frac{E_v - E_F}{k_b T}) = p = n_i,$$

$$\implies \boxed{n_i^2 = N_c N_v F_{\frac{1}{2}}(\frac{E_F - E_c}{k_b T}) F_{\frac{1}{2}}(\frac{E_v - E_F}{k_b T}),} \qquad (15.2)$$

which is a general relation for the intrinsic carrier concentration. Under non degenerate conditions, $E_F << E_c - k_b T$ and $E_F >> E_v + k_b T$. This is true when $E_v < E_F < E_c$, and the energy separation is several $k_b T$; in other words, E_F is deep in the bandgap of the semiconductor. In this situation, we get (see Fig. 15.2)

$$n \approx N_c e^{\frac{E_F - E_c}{k_b T}}, \text{ and } p = N_v e^{\frac{E_v - E_F}{k_b T}},$$

$$n = p = n_i \implies \boxed{n_i \approx \sqrt{N_c N_v} e^{-\frac{E_g}{2k_b T}}.} \qquad (15.3)$$

Thus, the 3D intrinsic carrier concentration goes as $n_i \sim T^{3/2} e^{-E_g/2k_b T}$, and is dependent only on the material bandstructure via the effective masses and the bandgap. As a simple example, if $m_c^\star = m_v^\star = 0.1 m_e$ and $E_g = 1$ eV, $n_i \approx 10^{10}/\text{cm}^3$ at $T = 300$ K. A narrow bandgap semiconductor has very high intrinsic carrier concentration, and a wide-bandgap semiconductor has very few, as is expected intuitively[1].

For GaAs, there is only one conduction band minimum with $m_c^\star \sim 0.067 m_e$, thus the conduction band valley degeneracy is $g_v^c = 1$. The valence band degeneracy is also $g_v^v = 1$ for each of the light-hole, heavy hole, and split-off bands. Because there are three distinct valence hands: the light hole (LH) band, the heavy hole (HH) band, and the split-off (SO) band (see Section 12.5), there will be holes in each of them. The total hole density is the sum of the hole densities of each band:

$$p = \underbrace{N_v^{LH} F_{\frac{1}{2}}(\frac{E_v^{LH} - E_F}{k_b T})}_{p_{LH}} + \underbrace{N_v^{HH} F_{\frac{1}{2}}(\frac{E_v^{HH} - E_F}{k_b T})}_{p_{HH}} + \underbrace{N_v^{SO} F_{\frac{1}{2}}(\frac{E_v^{SO} - E_F}{k_b T})}_{p_{SO}}, \qquad (15.4)$$

with the appropriate band-edges and effective masses. For example, if the LH and HH bands are degenerate at $\mathbf{k} = 0$ or the Γ-point of the

[1] This in turn has several implications on their use: it is rather difficult to make a narrow bandgap semiconductor electrically insulating at ambient temperatures. Famously, graphene has good electron transport properties, but since it has a zero bandgap, it cannot be made insulating: back in Exercise 5.2 you have calculated that the 2D carrier density varies as T^2, and cannot be lowered below $\sim 10^{11}/\text{cm}^2$ at room temperature. On the other hand, it is simpler to make a wide-bandgap semiconductor insulating, its conductivity can be varied over a *very* large range.

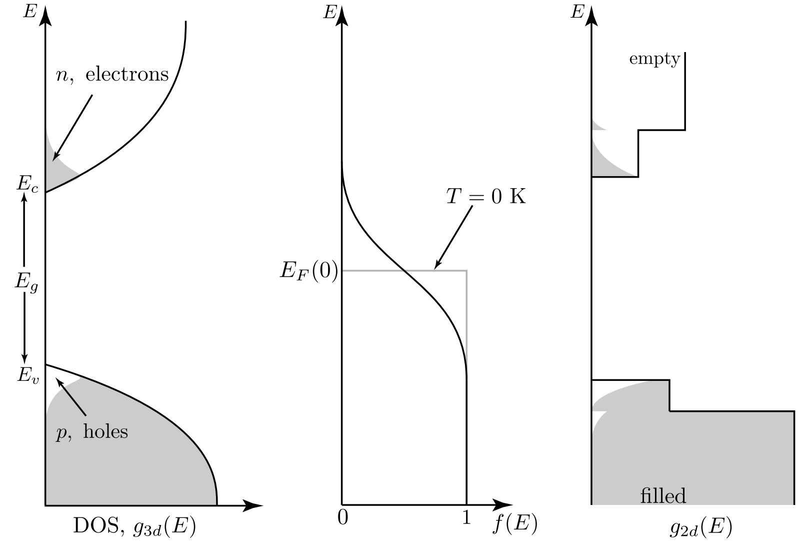


Fig. 15.2 Density of states $g(E)$, the Fermi function $f(E)$, and the occupied electron and hole densities for a 3D semiconductor, and a quantum well that hosts quasi-2D DOS. For intrinsic (undoped) semiconductors at thermal equilibrium, electrons are thermally ionized across the bandgap. Thus, the number of electrons in the conduction band n and the number of holes p in the valence band are the same, each equal to the intrinsic carrier concentration n_i characteristic of the semiconductor.

Brillouin zone $E_v = E_v^{LH} = E_v^{HH}$, and the split-off band is at energy Δ below E_0, the total hole density is given by

$$p = 2(\frac{2\pi k_b T}{h^2})^{\frac{3}{2}}[(m_v^{LH})^{\frac{3}{2}} + (m_v^{HH})^{\frac{3}{2}}]F_{\frac{1}{2}}(\frac{E_v - E_F}{k_b T}) + 2(\frac{2\pi m_v^{SO} k_b T}{h^2})^{\frac{3}{2}} F_{\frac{1}{2}}(\frac{E_v - \Delta - E_F}{k_b T}), \tag{15.5}$$

which may be written as

$$p = 2(\frac{2\pi m_v^{av} k_b T}{h^2})^{\frac{3}{2}} F_{\frac{1}{2}}(\frac{E_v - E_F}{k_b T}) + 2(\frac{2\pi m_v^{SO} k_b T}{h^2})^{\frac{3}{2}} F_{\frac{1}{2}}(\frac{E_v - \Delta - E_F}{k_b T}), \tag{15.6}$$

where $m_v^{av} = [(m_v^{LH})^{\frac{3}{2}} + (m_v^{HH})^{\frac{3}{2}}]^{\frac{2}{3}}$, which is sometimes[2] referred to as a "density of states effective mass".

Now for indirect bandgap semiconductors such as silicon and germanium, there are several equal-energy valleys in the conduction band, whose minima are not at the $\mathbf{k} = (0,0,0)$ or Γ- point. In silicon, there are $g_v = 6$ minima, each located $\sim 0.8\Gamma - X$ points, for example at $\mathbf{k}_{min} \approx (0.8\frac{2\pi}{a}, 0, 0)$, where a is the lattice constant. We label $E(\mathbf{k}_{min}) = E_c$ as the conduction band edge, for silicon $E_c - E_v = 1.1$ eV. The energy eigenvalues of the conduction bandstructure written out for small values[3] of $(k_x, k_y, k_z) = \mathbf{k}' - \mathbf{k}_{min}$ is given by

$$E(k_x, k_y, k_z) = E_c + \frac{\hbar^2}{2}(\frac{k_x^2}{m_1} + \frac{k_x^2}{m_2} + \frac{k_x^2}{m_3}), \tag{15.7}$$

where m_1, m_2, m_3 are the effective masses along the k_x, k_y, and k_z directions. For silicon, $m_1 = m_L \approx 0.98 m_e$ is the *longitudinal* electron effective mass for states moving in the six equivalent $\Gamma - X$ directions,

[2] The nomenclature is varied and sometimes confusing, so the emphasis should be in grasping the physical meaning.

[3] We are effectively shifting the origin of the $\mathbf{k}$-space to the conduction band minimum point, and keep the method general for all indirect bandgap semiconductors. For silicon, $(k_x, k_y, k_z) = \mathbf{k}' - \mathbf{k}_{min} = (k_x' - 0.8\frac{2\pi}{a}, k_y' - 0, k_z' - 0)$.

and $m_2 = m_3 = m_T \approx 0.19m_e$ is the electron *transverse* effective mass for carrier moving perpendicular to the $\Gamma - X$ directions, as indicated in Fig. 15.3. Now the electron density in the conduction band is written simply by summing over the *occupied* states in the **k**-space:

$$n = g_s g_v \sum_{\mathbf{k}} f(\mathbf{k}) = g_s g_v \int \frac{dk_x dk_y dk_z}{(2\pi)^3} \frac{1}{1 + e^{\frac{[E_c + \frac{\hbar^2}{2}(\frac{k_x^2}{m_1} + \frac{k_x^2}{m_2} + \frac{k_x^2}{m_3})] - E_F}{k_b T}}}, \tag{15.8}$$

which the transformation $k_x' = \sqrt{\frac{m_e}{m_1}} k_x$, $k_y' = \sqrt{\frac{m_e}{m_2}} k_y$, and $k_z' = \sqrt{\frac{m_e}{m_3}} k_x$ converts to a familiar form

$$n = g_s g_v \sqrt{\frac{m_1 m_2 m_3}{m_e^3}} \int \frac{dk_x' dk_y' dk_z'}{(2\pi)^3} \frac{1}{1 + e^{\frac{[E_c + \frac{\hbar^2}{2m_e}((k_x')^2 + (k_y')^2 + (k_z')^2)] - E_F}{k_b T}}}$$

$$= \underbrace{g_s g_v (\frac{2\pi m_{av} k_b T}{h^2})^{3/2}}_{N_c} F_{\frac{1}{2}}(\frac{E_F - E_c}{k_b T}), \text{ where } m_{av} = (m_1 m_2 m_3)^{\frac{1}{3}}. \tag{15.9}$$

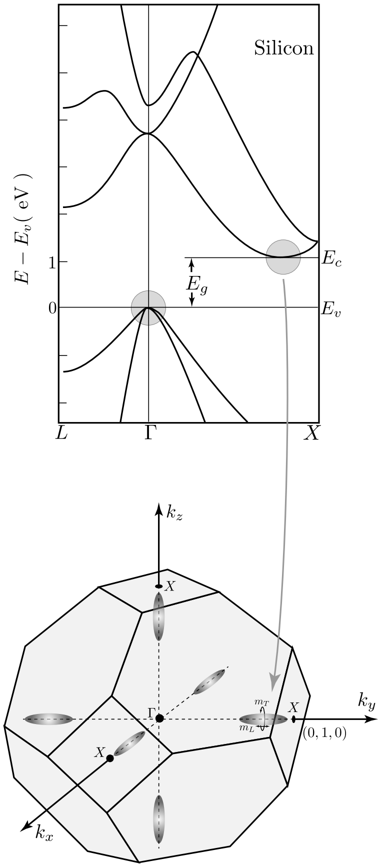


Fig. 15.3 Conduction band minima of silicon showing the $g_v = 6$ valleys along the six $\Gamma - X$ directions, and the origin of the DOS effective mass $(m_L m_T^2)^{\frac{1}{3}}$.

The resulting effective mass that enters N_c is $m_{av} = (m_1 m_2 m_3)^{\frac{1}{3}}$ and the appropriate valley degeneracy g_v must be used. For example, for silicon $g_v = 6$ for six equivalent $\mathbf{k}_{min}$ points along the $\Gamma - X$ axes, and $m_{av} = (m_L m_T^2)^{\frac{1}{3}}$. Similarly, for Ge the conduction band minima are in the $\Gamma - L$ directions at the BZ edge, but only half of each ellipsoid falls within the first BZ. This implies $g_v = \frac{1}{2} \cdot 8 = 4$ and $m_{av} = (m_L m_T^2)^{\frac{1}{3}}$, with $m_L = 1.6m_e$ and $m_T = 0.08m_e$. The carrier densities in the bands can be found in this manner for any sort of bandstructure. Because in an intrinsic semiconductor all electrons in the conduction band are thermally ionized from the valence band, their densities must be equal: $n = p$. This relation uniquely determines the Fermi level E_F. A more general relation for the Fermi level, which holds even when $n \neq p$ is obtained directly from Equation 15.3:

$$\boxed{E_F = \frac{1}{2}(E_c + E_v) + \frac{k_b T}{2} \ln(\frac{n}{N_c} \cdot \frac{N_v}{p}).} \tag{15.10}$$

In the intrinsic semiconductor, $n = p = n_i$, and if $N_c \approx N_v$, the Fermi level is located close to the middle of the bandgap. This is seen in Fig. 15.4. To be more accurate, the Fermi level is located closer to the band that has a *lower* DOS[4]. From Equation 15.10, if $n >> p$, the logarithmic term is positive and the Fermi level moves up from the mid-gap towards the conduction band edge E_c. If $p >> n$, the logarithmic term is negative and E_F moves down from the mid-gap towards E_v. This is shown schematically in Fig. 15.4, with the corresponding changes in the electron and hole concentrations in the bands. At thermal equilibrium, $np = n_i^2$ is still maintained, even when the Fermi level is not at the mid-gap energy, implying that even when say $n >> p$, we know *precisely* how few holes there are: it is $p = n_i^2/n$.

[4] The closer the Fermi level is to a band, the more *mobile* carriers in that band. Conversely, the Fermi level is closer to the band that has more mobile carriers.

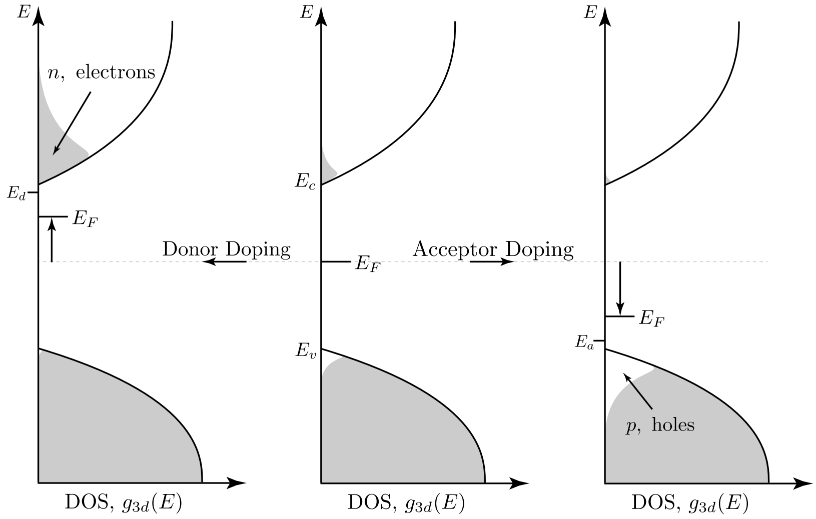


Fig. 15.4 Donor and acceptor doping in a 3D semiconductor. Doping with donor atoms moves the Fermi level E_F close to the conduction band, and the semiconductor becomes n-type. Doping with acceptors moves E_F close the valence band, making the semiconductor p-type.

Now the location of the Fermi level with respect to the band edges of the semiconductor is the **single most powerful tool** in the design of *any* semiconductor device, from diodes to transistors, from solar cells to LEDs and lasers. In the operation of each of these devices, the semiconductor is typically pulled out of thermal equilibrium by applying an external voltage, or due to incident photons. In this chapter, we discuss how we control the Fermi level of semiconductors at thermal equilibrium[5].

The Fermi level of semiconductors at thermal equilibrium is controlled by **chemical doping**. In Chapter 13 we discussed the atomic picture of doping. Replacing an atom of an intrinsic semiconductor with one that has one extra valence electron leads to n-type doping. This happens if the electron energy eigenvalue of the extra state E_d is located close to the conduction band edge, whereupon the extra electron instead of being tied to the donor atom is thermally ionized into the band and becomes mobile. This is shown in the energy band diagram picture in Fig. 15.5. For this to happen effectively at room temperature, the donor energy level should be within $\sim k_b T \approx 26$ meV below the conduction band edge. The exact complement of this form of n-doping is if the dopant atom has one less valence electron than the atom it substitutes: this acceptor dopant captures an electron from the filled valence band, thus populating the filled valence band with a mobile hole. The ionized donor atom is therefore +vely charged, and the ionized acceptor atom is -vely charged, as seen in Fig. 15.5. The Fermi level moves closer to the conduction band for donor doping, and closer to the valence band for acceptor doping, seen in Fig. 15.4.

How do we now find the Fermi level for a general doping situation? The answer to this question is always the same: **To find the Fermi level, write down the equation for charge neutrality. The Fermi level is located at the energy that exactly balances all negative and all**

---

[5] Under non-equilibrium situations, the relation $np = n_i^2$ changes to the form $np = n_i^2 \cdot \exp \frac{E_{Fn} - E_{Fp}}{k_b T}$, where E_{Fn} and E_{Fp} are the electron and hole **quasi Fermi levels**. We will discuss and use this concept in later chapters when we discuss semiconductor devices.

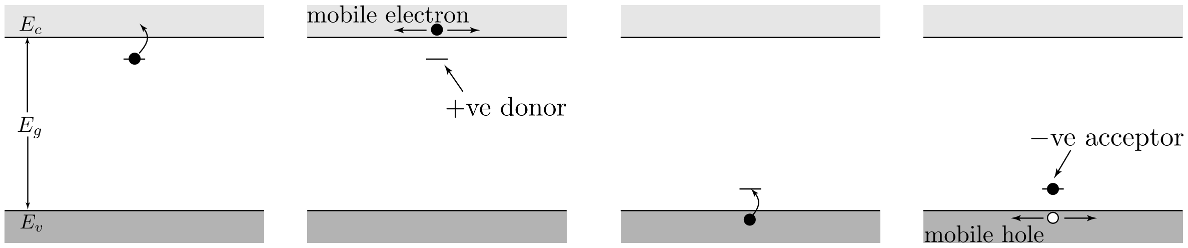


Fig. 15.5 Donor doping populates the conduction band with mobile electrons, leaving behind a +ve bound donor ion. Acceptor doping populates the valence band with mobile holes, leaving behind a -ve bound acceptor ion.

positive charges. Compared to the mobile electrons n and mobile holes p that were the only charges in the intrinsic semiconductor, in a doped semiconductor we must now also account for the +vely charged donors and -vely charged acceptors. Out of N_D density of donors, a fraction N_D^+ are ionized; similarly for N_A acceptors we have N_A^- ionized acceptors. The charge neutrality equation now reads

$$n + N_A^- = p + N_D^+. \tag{15.11}$$

In this charge neutrality equation, choosing a semiconductor and its doping densities N_D and N_A fixes all terms of the equation, leading to a unique determination of E_F. To find it, we must first know what fraction of dopants are ionized: i.e., what are N_D^+/N_D and N_A^-/N_A?

It is enticing to use the Fermi–Dirac distribution directly to find the occupation functions of the donor states E_d and acceptor states E_a, but alas, that is incorrect, as has been found experimentally. The reason is rather interesting: here we encounter a taste of a genuine *many-particle effect* which makes the Fermi–Dirac distribution invalid. Consider a donor state of energy E_d as shown in Fig. 15.6. Here are all the possible scenarios for this energy state: the extra electron it brought in may be ionized into the band, in which case the state itself is occupied with $n = 0$ electrons, with total energy 0. The electron may not be ionized, in which case the state is filled with $n = 1$ electron, with either up or down spin, with energy E_d for each case.

A third possibility is an intriguing one: since the donor eigenvalue can hold two electrons of opposite spins, this forms another state, but with a twist. Recall from Chapter 14 that the electron envelope wavefunction $|C(r)|^2 \sim e^{-r/a_B^\star}$ of the donor-bound electron state is *not* an extended Bloch function, but *localized* like a Hydrogen atom with an effective Bohr radius $a_B^\star$ that extends over $\sim$ 10s of lattice constants. What this means is that the second electron is physically at roughly a distance $a_B^\star$ from the first electron, and must experience a strong Coulomb repulsion. This repulsion energy U thus raises the total energy of the two electron state to[6] $2E_D + U$, where U is typically in the eV range, much larger than k_bT. This strong electron-electron interaction violates the intrinsic assumption of non-interacting electrons used to derive the Fermi Dirac distribution in Chapter 4, Section 4.4. Thus,

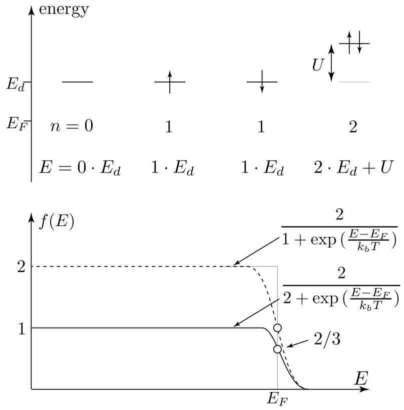


Fig. 15.6 On-site Coulomb repulsion, an electron-electron (or many-particle) interaction effect modifies the occupation statistics of dopants, making it different from the Fermi–Dirac distribution.

---

[6] Since a single electron feels no Coulomb repulsion due to itself, the factor U enters only if there are two electrons. This is the reason the energy is *not* $2(E_d + U)$.

we must re-derive the many-particle distribution function. It is as simple as the original derivation, and the result looks deceptively similar, but is not quite the same.

The derivation of the distribution function uses the same recipe as used in Section 4.4: find all system energy states, find[7] the partition function $Z = \sum_n e^{-\beta(E_n - nE_F)}$, and then the probability of each state n_i is $e^{-\beta(E_{n_i} - n_i E_F)}/Z$. The values of n and the corresponding E_n are shown in Fig. 15.6. The average occupation function of the state is:

$$\begin{aligned}\langle n \rangle = \frac{0 \cdot e^0 + 1 \cdot e^{-\beta(E_d - E_F)} + 1 \cdot e^{-\beta(E_d - E_F)} + 2 \cdot e^{-\beta(2E_d + U - 2E_F)}}{e^0 + e^{-\beta(E_d - E_F)} + e^{-\beta(E_d - E_F)} + e^{-\beta(2E_d + U - 2E_F)}} \\ = \frac{1 + e^{-\beta(E_d + U - E_F)}}{1 + \frac{1}{2}e^{-\beta(E_d - E_F)} + \frac{1}{2}e^{-\beta(E_d + U - E_F)}} \\ \underbrace{\approx}_{U \to 0} \frac{2}{1 + e^{\beta(E_d - E_F)}} \\ \underbrace{\approx}_{U \to \infty} \frac{2}{2 + e^{\beta(E_d - E_F)}}.\end{aligned} \tag{15.12}$$

If the electron-electron Coulomb interaction energy is negligible, by taking the limit $U \to 0$, the average occupation function is $\langle n \rangle = 2f(E_d)$, where $f(E)$ is the Fermi–Dirac distribution function. We expect this, because each eigenvalue can hold two electrons of opposite spin, the 2 here is just the spin degeneracy. In the opposite limit of strong electron-electron interaction, the occupation number of the state at thermal equilibrium is not a Fermi–Dirac distribution: this is easily seen by putting $E_d = E_F$, which yields an unconventional occupation function[8] of $2/3$. For the non-interacting distribution, this number is 1, exactly half of the 2 maximum electrons. These two distribution functions are shown in Fig. 15.6.

Since the donor atom is ionized and contributes an electron to the conduction band if its occupation is zero (it has lost its electron!), we obtain immediately the many-particle result of the ionized donor density, (and the ionized acceptor density based on identical arguments[9]):

$$\frac{N_D^+}{N_D} = \frac{1}{1 + 2 \cdot e^{\frac{E_F - E_d}{k_b T}}}, \quad \frac{N_A^-}{N_A} = \frac{1}{1 + 2 \cdot e^{\frac{E_a - E_F}{k_b T}}}, \tag{15.13}$$

which immediately leads to the master charge-neutrality equation

$$\boxed{\underbrace{N_v F_{\frac{1}{2}}(\frac{E_v - E_F}{k_b T})}_{p} + \underbrace{\frac{N_D}{1 + 2 \cdot e^{\frac{E_F - E_d}{k_b T}}}}_{N_D^+} = \underbrace{N_c F_{\frac{1}{2}}(\frac{E_F - E_c}{k_b T})}_{n} + \underbrace{\frac{N_A}{1 + 2 \cdot e^{\frac{E_a - E_F}{k_b T}}}}_{N_A^-}.} \tag{15.14}$$

---

[7] Here $\beta = 1/k_b T$.

[8] This is seen experimentally in what is known as the 2/3 conductance quantization peak in ballistic transport, for the same many-particle interaction reason as discussed here.

[9] For acceptor doping, the prefactor of 2 in the exponent in the denominator depends on the degeneracy of the valence band edge. If the LH and HH bands are degenerate at the Γ-point, then this factor is 4, and the ionization is $\frac{N_A^-}{N_A} = \frac{1}{1+4 \cdot e^{\frac{E_a - E_F}{k_b T}}}$.

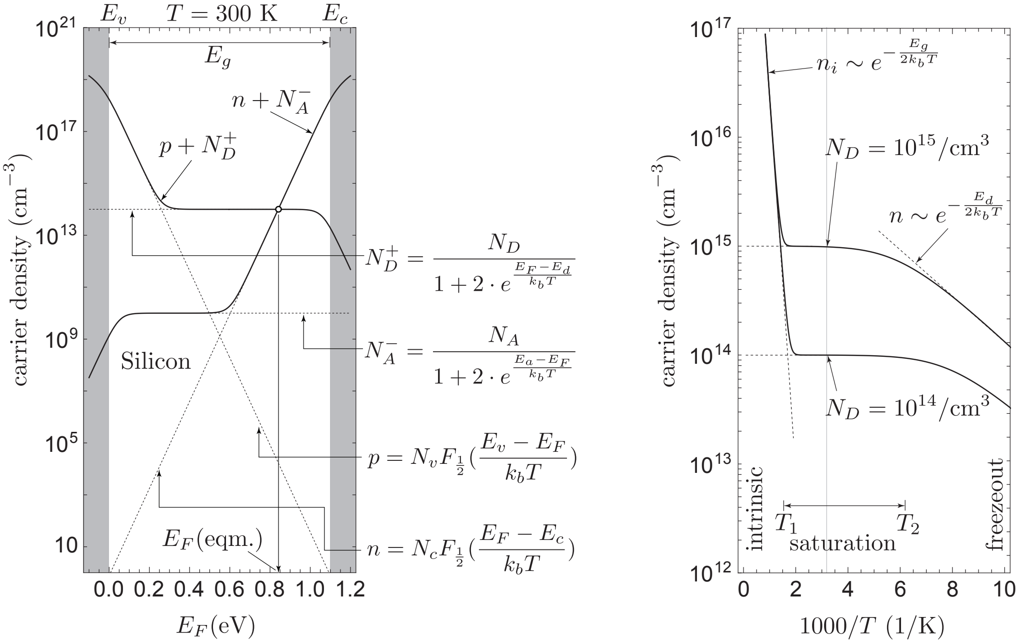


Fig. 15.7 Carrier statistics in silicon.

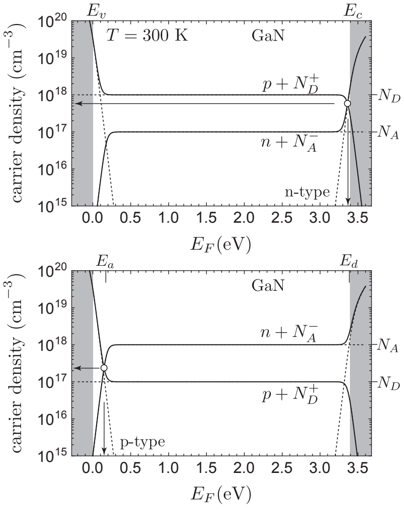


Fig. 15.8 Carrier statistics in doped GaN: n-type and p-type. Note that because the acceptor ionization energy is larger than the donor ionization energy, the p-type hole density is lower than the n-type electron density for the same effective doping.

The solution of this equation yields the Fermi level E_F at thermal equilibrium. Once the location of E_F is known, the mobile carrier densities n, p and the ionized donor and acceptor densities N_D^+ and N_A^- are all known, and the state of the semiconductor is uniquely determined. Fig. 15.7 shows a graphical solution to the charge neutrality equation for silicon, with doping densities $N_D = 10^{14}/\text{cm}^3$ and $N_A = 10^{10}/\text{cm}^3$. The y-axis is the density, and the x-axis is the location of the Fermi level E_F with respect to the valence band edge E_V. Each term of the equation are indicated as dashed lines in the plot, and the solid lines are the left-hand side (LHS) and the right-hand side (RHS). They intersect slightly above $E_F \approx 0.8$ eV, which is the location of the Fermi level where LHS = RHS, and the charges are completely balanced. The corresponding electron density can be read off as $n \approx N_D = 10^{14}/\text{cm}^3$, and the hole density is $p \approx n_i^2/N_D$, below $10^5/\text{cm}^3$. The point of intersection is along the line $n = N_c F_{1/2}[(E_F - E_c)/k_b T]$, in the part that is still exponential, in the Maxwell-Boltzmann approximation $n \approx N_C \exp[-(E_c - E_F)/k_b T]$. If the doping N_D was much higher, the point of intersection will be closer to the conduction band edge $E_F \approx E_c$, where the n-line starts curving slightly, which is the degenerate doping limit. If the acceptor density was higher than the donor density, the intersection point would have been on the left, on the line of p instead, making the semiconductor p-type.

Fig. 15.8 shows the charge neutrality plot for GaN, which has a much larger bandgap than silicon. Because the acceptor ionization energy of GaN doped with Mg $E_a \sim 180$ meV is deeper than the donor ionization energy $E_d \sim 20$ meV for GaN doped with Si, the hole density for the same acceptor doping is lower. Fig. 15.9 shows how the Fermi level moves as a function of temperature, for various donor and acceptor doping densities for silicon, and GaN. At high temperatures, the Fermi level moves towards the middle of the gap, because the thermally generated interband carrier density increases. A wider bandgap semiconductor is more resistant to this movement of the Fermi level.

If the doping densities are low, $N_D << N_c$ or $N_A << N_v$, the Fermi–Dirac integrals of Equation 15.14 may be approximated by their Maxwell–Boltzmann form:

$$N_v e^{\frac{E_v-E_F}{k_bT}} + \frac{N_D}{1+2\cdot e^{\frac{E_F-E_d}{k_bT}}} = N_c e^{-\frac{E_c-E_F}{k_bT}} + \frac{N_A}{1+2\cdot e^{\frac{E_a-E_F}{k_bT}}}. \tag{15.15}$$

Typical semiconductor layers are doped either with donors (for n-type), or with acceptors (for p-type). For a donor-doped semiconductor of doping N_D, assuming $N_A = 0$, this equation further simplifies to

$$N_v e^{\frac{E_v-E_F}{k_bT}} + \frac{N_D}{1+2\cdot e^{\frac{E_F-E_d}{k_bT}}} = N_c e^{-\frac{E_c-E_F}{k_bT}}. \tag{15.16}$$

The numerical solution[10] of the carrier density n for this n-doped semiconductor for silicon is shown in the right panel of Fig. 15.7 for doping densities $N_D = 10^{14}/\text{cm}^3$ and $N_D = 10^{15}/\text{cm}^3$ as a function of inverse temperature. Over a temperature range $T_2 < T < T_1$ (note that temperature increases to the left and decreases to the right), the mobile electron density is found to be equal to the doping density $n \approx N_D$. This is called the *saturation density regime* of doping, when (almost) all dopants are ionized. This window typically includes 300 K, or room temperature as indicated by the vertical gray line.

At temperatures higher than T_1 (to the left of the figure), the electron density increases exponentially as $n_i \sim \sqrt{N_c N_v} e^{-E_g/2k_bT}$, and is called the *intrinsic density regime*. The meaning is clear: N_D donor atoms can provide a maximum of N_D electrons when they are fully ionized. But the valence band has a far larger store of electrons, equal to the density of atoms – which can be thermally ionized across the bandgap into the conduction band. For temperatures lower than T_1, the donors win out, but for higher temperatures $T > T_1$, interband thermal generation overwhelms the contribution from donors.

In the other end of the spectrum, at low temperatures $T < T_2$ to the right of the figure, interband generation is very weak, and the electrons that at room temperature were thermally ionized fall back to the respective donors. This is aptly called the *carrier freeze-out regime*. In this regime, the carrier density depends on temperature as $n \approx \sqrt{N_c N_d/2} \cdot e^{-E_d/2k_bT}$, and is used as an experimental method to directly measure the donor activation energy E_d. The intrinsic regime has an activation energy half the semiconductor bandgap $E_g/2$, whereas the activation energy for the freezeout regime is half the dopant activation energy $E_d/2$.

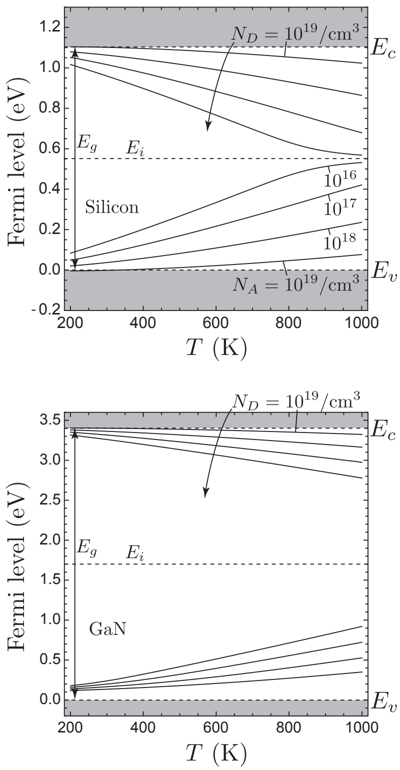


Fig. 15.9 Location of the Fermi level E_F in silicon and GaN as a function of temperature for various donor and acceptor doping densities.

---

[10] This equation can be analytically solved for E_F, it can be seen to be *cubic* in e^{E_F/k_bT}.

15.2 E_F is constant at thermal equilibrium

In the past section we have found how the technique of doping places the Fermi level of a semiconductor at a desired position in energy: closer to the conduction band edge for n-type mobile carriers, or closer to the valence band edge for p-type. A semiconductor device will typically have junctions of semiconductors of different doping in different regions in space, or junctions of metals and semiconductors. To understand all junctions, we derive what may be stated unequivocally as the *central tenet of semiconductor device physics*:

$$\boxed{\textbf{At thermal equilibrium, the Fermi level } E_F \textbf{ is constant in space.}} \tag{15.17}$$

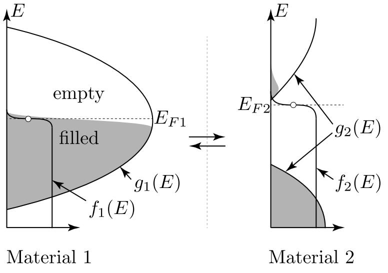


Fig. 15.10 Figure illustrating why at thermal equilibrium, the Fermi level E_F must be equal everywhere.

To derive this result, Fig. 15.10 shows a metal (Material 1) and a n-type doped semiconductor (Material 2) before the junction is formed. It is emphasized that the following argument is a powerful one: it holds for two metals, metal/semiconductor, n- and p-semiconductor, semiconductors of different bandgaps, or even their junctions with topological insulators. The two materials have their individual DOS $g_1(E)$ and $g_2(E)$. Before the two materials form a junction, they have electrons filling the DOS at energy E with probability given by the Fermi–Dirac distribution $f(E) = 1/[1+\exp\left(\frac{E-E_F}{k_bT}\right)]$ as $f_1(E)$ and $f_2(E)$ according to two Fermi levels E_{F1} and E_{F2}.

When the junction is formed, electrons can move between the two materials. The rate of transfer of electrons at any energy E from material 1 to material 2 must be proportional to the number of filled states n_1 in material 1 and number of empty states p_2 in material 2 at that energy. Note that this is a direct consequence of the Pauli exclusion principle, because electrons are fermions. Thus, we must have

$$\underbrace{g_1(E)\cdot f_1(E)}_{n_1}\cdot\underbrace{g_2(E)\cdot(1-f_2(E))}_{p_2} = \underbrace{g_2(E)\cdot f_2(E)}_{n_2}\cdot\underbrace{g_1(E)\cdot(1-f_1(E))}_{p_1}$$
$$\implies f_1(E) = f_2(E) \implies \boxed{E_{F1} = E_{F2}.} \tag{15.18}$$

This completes the proof of statement 15.17 that E_F is constant everywhere, in any combination of junctions at thermal equilibrium[11].

15.3 Metal-semiconductor Schottky junctions

Most semiconductor devices have regions of different doping, and also of different physical properties next to each other. These include metal-semiconductor (Schottky) junctions, junctions between semiconductor regions doped differently in space (e.g. p-n junctions), and heterojunctions of semiconductors of different bandgaps and doping. Let us extend our knowledge of the effect of doping and junctions by considering such junctions. We will make our way towards being able to draw the *energy band diagrams* of these junctions. *While in energy bandstructure we plot the energy eigenvalues vs. k as bands $E_n(k)$, in the energy band diagram we plot only the conduction band minimum E_c, the valence band maximum E_v, and the bandgap E_g as a function of the space coordinate.* Thus, energy band diagrams vividly show the propensity of electrons and holes to move in the device, as electrons roll down the conduction band edge, whereas holes bubble up the valence band edge. Being able to draw the energy band diagram thus takes central importance in semiconductor device design and analysis (see Exercise 15.4).

---

[11] It is a strong statement: it holds at any temperature, and for any DOS across the junction. If you cascade several junctions, it still holds: the Fermi level is flat throughout any complicated structure with multiple junctions one can cook up! It is very difficult to wiggle one's way out of the leveling power of thermodynamics at equilibrium! Only when the junctions are pulled out of equilibrium is the rate of electron transfer between 1 ↔ 2 unequal: and that is precisely the current flow, which will be the subject of Module III of this book. In this chapter, we explore the consequences of the fact that E_F is constant for a few junctions at equilibrium.

Fig. 15.11 shows a junction formed between a metal and a n-type doped semiconductor. The top panel of the figure shows the metal Fermi level E_{Fm}, which is at an energy $q\phi_m$ below the vacuum level: this is the workfunction of the metal. The vacuum level to the conduction band edge of the semiconductor is characteristic of the semiconductor material, and is known as the electron affinity, denoted by $q\chi_s$. The valence band edge is thus $q\chi_s + E_g$ below the vacuum energy. If the semiconductor is doped with donors and we are in the saturation regime, the Fermi level is closer to the conduction band edge than the valence band, determined by $n \approx N_d \implies E_{Fs} \approx E_c - k_b T \ln(N_c/N_d)$. This is the situation *before* the junction is formed.

When the junction is formed, electrons can flow from one material to another. One may be inclined to think that a metal has far more mobile electrons $\sim 10^{23}/\text{cm}^3$ than the semiconductor (say $\sim 10^{17}/\text{cm}^3$), and therefore electrons should flow from the metal to the semiconductor. However, this would be incorrect for the junction shown in Fig. 15.11! Which way electrons will flow is determined by the location of the Fermi levels. **Electrons flow from a region of higher Fermi level to a region with a lower Fermi level, till the Fermi level is equal everywhere: this is how equilibrium is achieved**. We proved this assertion in the last section. Since in the metal-semiconductor junction considered here, the Fermi level of the semiconductor E_{Fs} is higher than the metal Fermi level E_{Fm}, the semiconductor will lose some electrons to the metal. The voltage corresponding to the difference in the Fermi levels $qV_{bi} = E_{Fs} - E_{Fm}$ is called the **built-in potential** of the junction. How many electrons will the semiconductor lose to the metal? This question has a precise answer, and forms the genesis of the energy band diagram.

When the n-type doped semiconductor loses electrons, it becomes positively charged. The positive charges are the bound donor atoms from which the electrons came in the first place – this process is therefore called **depletion**. The semiconductor is thus depleted of mobile carriers at the metal-semiconductor junction, but how deep x_d does this depletion region go into the semiconductor? As the semiconductor becomes positively charged, electric field lines start at the positive ionized donor atoms, and terminate on the excess electrons in the negatively charged metal. Since the metal allows no electric fields to exist in the interior in DC conditions, the excess electrons are on the surface of the metal, at the junction. Back in Chapter 2, Exercise 2.4, you

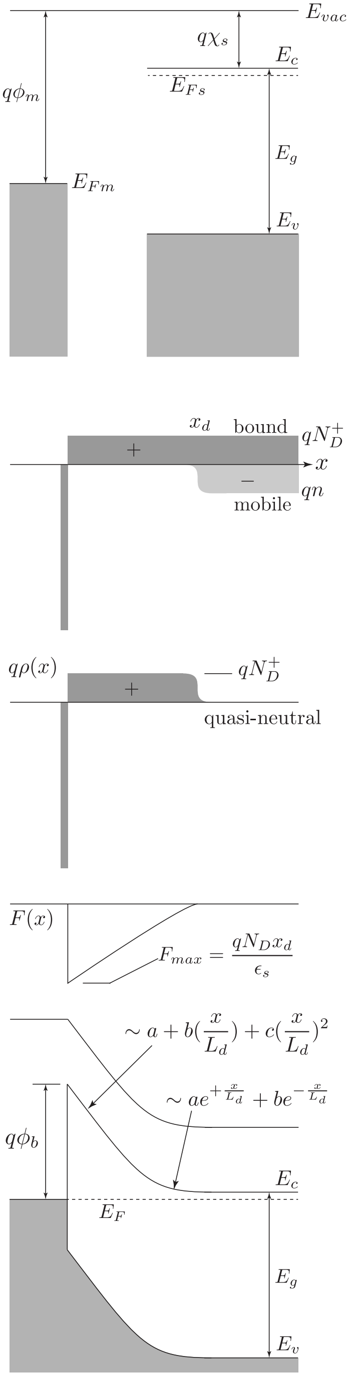


Fig. 15.11 Electrostatics of the metal-semiconductor Schottky junction. To draw the energy band diagram shown at the bottom, a good strategy is to take the path [charge → field → bands].

had found that by Gauss's law, the electric field between two equal and opposite 2D sheet charge densities σ is equal to $F = \sigma/\epsilon_s$, where ϵ_s is the dielectric constant of the semiconductor. If the thickness of the depletion region is x_d, then the total positive *sheet* charge density is $\sigma_s = +qN_D x_d$. The metal has an exactly equal and opposite sheet charge of density $\sigma_m = -\sigma_s$. By Gauss's law, the electric field at the metal/semiconductor junction is therefore $F_{max} = qN_D x_d/\epsilon_s$. If the depletion region has uniform doping, the charge N_D^+ is constant near the junction. The Poisson equation relates the electric field to the charge: $\frac{dF(x)}{dx} = \frac{qN_D^+}{\epsilon_s}$, which implies that near the junction, the electric field changes linearly with distance x into the semiconductor. This is shown in Fig. 15.11. The area under the electric field curve $-\int_{x=0}^{x=x_d} F(x)dx = \frac{1}{2}F_{max}x_d = \frac{qN_D x_d^2}{2\epsilon_s}$ must equal the potential difference, which is exactly the built-in voltage V_{bi}. Since V_{bi} is locked by the choice of semiconductor, doping, and metal, we get the depletion region thickness must be

$$\boxed{\frac{E_{Fs} - E_{Fm}}{q} = V_{bi} = \frac{qN_D x_d^2}{2\epsilon_s}} \implies \boxed{x_d = \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}}.} \qquad (15.19)$$

For example, if $V_{bi} = 0.5$ eV and $N_D = 10^{16}/\text{cm}^3$ for a semiconductor of dielectric constant $\epsilon_s = 10\epsilon_0$, the depletion region thickness is $x_d \approx 0.24\mu$m, or ~ 235 nm.

We have implicitly made the assumption that the edge of the depletion region is sharply defined. This is not exactly correct, but is an excellent approximation. To solve for the electric field profile and the band profile exactly, we must solve the Poisson equation

$$-\frac{d^2V(x)}{dx^2} = \frac{\rho(x)}{\epsilon_s} = \frac{q}{\epsilon_s}[p(x) + N_D^+(x) - n(x) - N_A^-(x)]. \qquad (15.20)$$

To solve this equation, we realize that when in the neutral regions of the semiconductor as $x \to \infty$, the conduction band electron density $n = N_D$, and because $n(x) = N_c e^{-(E_c(x)-E_F)/k_bT}$, in a depletion region the carrier density must decrease exponentially as the conduction band moves away from the Fermi level. We denote the amount of *band bending* $\psi(x) = [E_c(x) - E_c(x \to \infty)]/k_bT$ as a non-dimensional parameter, which lets us write $n(x) = N_D e^{-\psi(x)}$, with a transparent physical meaning. Multiplying Equation 15.20 throughout by $-q/k_bT$ and identifying that $\psi(x) = qV(x)/k_bT$ is also the same non-dimensional potential-drop term, the Poisson equation becomes[12]

$$\frac{d^2\psi}{dx^2} = \frac{q^2N_D}{\epsilon_s k_b T}[1 - e^{-\psi} + \frac{n_i^2}{N_D^2}e^{\psi}] \approx \frac{1}{L_D^2}[1 - e^{-\psi}], \qquad (15.21)$$

where we identify $L_D = \sqrt{\epsilon_s k_b T/q^2 N_D}$, called the **Debye screening length**, is the natural length scale over which the potential can change. For $\psi >> 1$, $e^{-\psi} \approx 0$, which is the case near the metal/semiconductor

[12] This form of solving the Poisson equation is used to understand the physics of a field-effect transistors, p-n diodes, bipolar transistors, and heterojunctions.

junction and most of the depletion region except near $x \approx x_d$. Then the solution is a quadratic function

$$\frac{d^2\psi}{dx^2} \approx \frac{1}{L_D^2} \implies \psi(x) = a + b(\frac{x}{L_D}) + c(\frac{x}{L_D})^2 \tag{15.22}$$

with the constants a, b, c determined by the boundary conditions. For the depletion region edge near $x = x_d$, $\psi << 1$, the exponential is Taylor expanded as $e^{-\psi} \approx 1 - \psi$ to get

$$\frac{d^2\psi}{dx^2} \approx \frac{1}{L_D^2}[1 - e^{-\psi}] \approx \frac{\psi}{L_D^2} \implies \psi(x) = ae^{\frac{x}{L_D}} + be^{-\frac{x}{L_D}}. \tag{15.23}$$

The lowest panel in Fig. 15.11 shows the energy band diagram of the resulting Schottky junction. Near the surface, there is a large amount of band-bending, $\psi >> 1$, and the band $E_c(x)$ bends in a parabolic fashion. Near the edge of the depletion region, the band-bending $\psi << 1$ is small, and the band bends in an exponential fashion. Typically this change is neglected and the assumption of a sharp depletion edge is made to simplify the design and analysis, but numerical design programs can take these features into account.

The total band-bending in the semiconductor can be used now to determine the coefficients a, b, c in the solution of the Poisson equation. They can also be written down by inspection: the total band bending is identified as exactly equal to the built-in potential $\psi(0)k_bT = qV_{bi} = E_{Fs} - E_{Fm} = q\phi_m - (q\chi_s + k_bT\ln[N_c/N_d])$. By energy conservation[13], the height of the barrier formed for electrons to move from the bottom of the conduction band in the neutral region of the semiconductor to the metal is equal to $\psi(0)k_bT = q\phi_m - q\chi_s$. The barrier for electrons to move from the metal to the semiconductor $q\phi_b = q\phi_m - q\chi_s$ is slightly higher, and is called the **Schottky barrier height**.

15.4 p-n homojunctions

Fig. 15.12 shows the junction formed between n-type and p-type doped regions of the *same* semiconductor. The bandgap is the same for both sides, but the Fermi level E_{Fn} is close to the conduction band edge in the n-side, and E_{Fp} is close to the valence band edge on the p-side. Since the location of the Fermi levels depend on the doping densities N_d and N_a, we can immediately find the built-in voltage. Because $n \approx N_D = N_c e^{-[E_c - E_{Fn}]/k_bT}$, we have $E_{F_n} = E_c - k_bT\ln[N_c/N_D]$. Similarly, for the p-side, $E_{F_p} = E_v + k_bT\ln[N_v/N_A]$. Using the dependence of the intrinsic carrier concentration $n_i^2 = N_cN_ve^{-E_g/k_bT}$, we get the built-in voltage as

$$\boxed{qV_{bi} = E_{Fn} - E_{Fp} = E_g - k_bT\ln[\frac{N_AN_D}{N_cN_v}] = k_bT\ln[\frac{N_AN_D}{n_i^2}].} \tag{15.24}$$

---

[13] In Chapter 18 we explore how this energy band diagram and the potential barrier explains the rectifying behavior, its behavior as a variable capacitor, as a gate of transistors, for low-resistance tunneling ohmic contacts with heavy doping, and several related useful properties of the Schottky diode, making it a useful semiconductor device since the earliest days of the field.

From what we learned about the Schottky diode, we can already say that the barrier for mobile carriers to move from one side of the junction to the other must be equal to qV_{bi}. If the n and p sides are doped (heavily) at densities $N_D \approx N_c$ and $N_A \approx N_v$, the built-in voltage is equal to the bandgap $qV_{bi} \approx E_g$ of the semiconductor. Typical doping densities are lower, resulting in a V_{bi} that is typically smaller than the bandgap by several $k_b T$. The charge transfer that occurs near the p-n junction is similar in spirit to the Schottky diode, but there is an important difference. Unlike the metal, a semiconductor allows electric fields to exist inside, by depleting mobile carriers, and exposing the bound dopant ions. So the electrons that flow from the n-side fills the hole states in the valence band of the p-side. Since the hole density $p \approx N_A^-$ is equal to the acceptor doping density far from the junction, near the junction the net unbalanced charge density becomes N_A^-. The n-side of the story is similar to the Schottky diode. This sets up a bound charge dipole between positively charged ionized donors of density N_D^+ over a depletion thickness x_n, and of negatively charged ionized acceptors of charge density N_A^- to a depth of x_p from the junction. Charge neutrality requires $N_D^+ x_n = N_A^- x_p$. The net unbalanced charge $q\rho(x)$ is shown as a dipole in Fig. 15.12.

The corresponding electric field profile $F(x)$ still peaks at the junction, but extends into both sides. The maximum field is now connected to both the depletion thicknesses via $F_{max} = qN_D x_n/\epsilon_s = qN_a x_p/\epsilon_s$. Since the area under the electric field triangle profile is equal to the built-in voltage drop across the junction, using the charge neutrality relation $N_D^+ x_n = N_A^- x_p$ we find

$$V_{bi} = \frac{1}{2} F_{max}(x_n + x_p) \implies F_{max} = \sqrt{\frac{2qN_A N_D V_{bi}}{\epsilon_s (N_A + N_D)}},$$

$$x_n + x_p = W = \sqrt{\frac{2\epsilon_s}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right) \cdot V_{bi}}, \text{ and}$$

$$x_n = \frac{N_A}{N_A + N_D} W, \text{ and } x_p = \frac{N_D}{N_A + N_D} W. \tag{15.25}$$

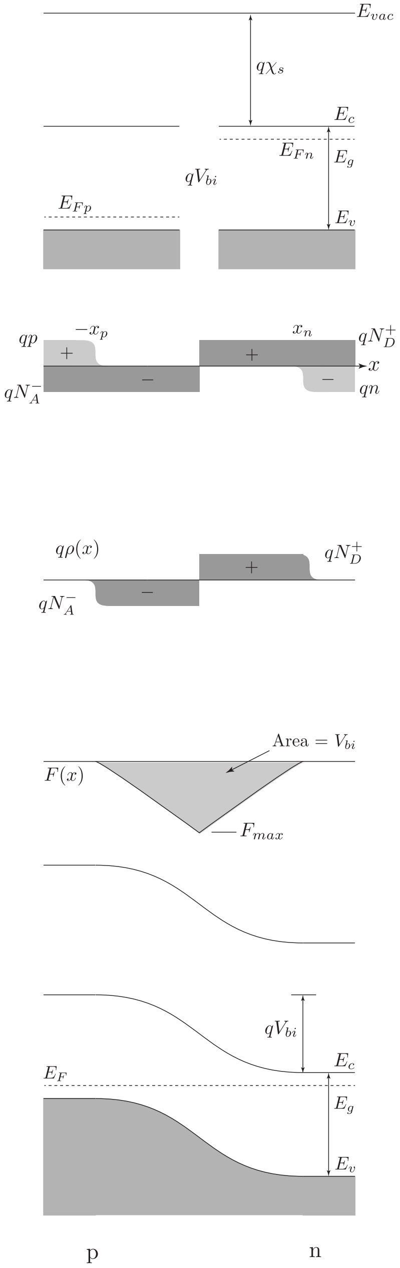


Fig. 15.12 Electrostatics of a p-n junction diode.

We note that the depletion region is divided between the two sides in the *inverse ratio* of the doping densities: the depletion region is thicker in the side that is doped lighter. For example, if a pn junction is formed with $N_A >> N_D$, the depletion region thickness is $W \approx \sqrt{2\epsilon_s V_{bi}/qN_D}$. We have ended up with the same depletion thickness as the Schottky diode (Equation 15.19)! This should not be surprising, because a heavily doped semiconductor is in many aspects similar to a metal: it can provide a lot of bound charge within a very short depletion thickness, and therefore does not let the electric field penetrate too deep. On the other hand a lightly doped semiconductor must deplete great distances to terminate, or provide electric field lines. This asymmetry of the depletion thicknesses and fields in a pn junction is a very useful feature that is exploited rather heavily in devices: making a junction one-sided is responsible for unidirectionality and *gain* in electronics and photonics as discussed in later chapters. Finally, we note that if the doping is high on both sides of the junction, the depletion thickness is very narrow: this feature is exploited to make interband tunneling diodes, as will be discussed in Chapter 24.

The energy band diagram is now obtained by integrating the electric field profile. From the discussion on Schottky diodes, it is evident that near the junction, the band bending is large, $\psi(x) >> 1$ in the n-side (or the p-side). This is clear since the band bending is of the order of the bandgap, which is much larger than $k_b T$. Thus, the band bending is parabolic near the junction, with the curvature of the band edges proportional to the ionized dopant densities on each side. Since the depletion region in the p-side has negative charge N_A^-, the curvature of the bands points downwards there, and in the n-side the +ve N_D^+ creates an upward band curvature. This is a general feature of energy band diagrams: note that the positively charged donors of the metal/n-semicondudctor Schottky diode in Fig. 15.11 also create a positive curvature.

From the energy band diagram of the pn homojunction in Fig. 15.12, we note that electrons in the conduction band on the n-side must surmount a potential energy barrier of height qV_{bi} to move to the p-side. Since holes bubble up the bands, the holes in the valence band on the p-side face the exact equal potential barrier to move to the n-side. In later chapters, we will see how applying a voltage will either lower, or increase the barrier heights, injecting electrons through the depletion region into the p-side, where they will become *minority carriers* because their number is lower than holes in the p-side. Similarly, holes will be injected into the n-side, where they are minority carriers. For this reason, a pn junction is called a minority-carrier device. If we play our cards right, we can make electrons and holes recombine to produce light: this is the working principle of semiconductor light-emitting diodes, and lasers, discussed in later chapters of this book.

15.5 Heterojunctions

In the homojunction, the potential barrier for electrons is the same as for holes. For several device applications, a very powerful tool can introduce *unequal* barrier heights for electrons and holes: this is achieved by using heterostructures. In Chapter 14 we discussed that semiconductors of unequal bandgaps can be realized in various quantized structures. The top three panels of Fig. 15.13 show the three possible band alignments at the heterojunction formed between two undoped semiconductors. These band alignments are possible by choosing appropriate semiconductors and growing them epitaxially. The conduction band offsets ΔE_c and the valence band offsets ΔE_v are locked by the choice of the semiconductors. Fig. 15.14 shows a range of bandgaps and band alignments available with 3D semiconductors, and Fig. 15.14 shows the 2D semiconductor counterpart. For 3D semiconductors, epitaxial growth with low defect densities requires lattice-matching, or strained layers up to certain critical thicknesses can be realized as long as the semiconductors are almost lattice matched. For 2D materials that have weak van der Waals interlayer bonding, this stringent condition of lattice matching is partially relaxed.

Among the alignments shown in Fig. 15.13, the **Type I**, or straddling band alignment is most common: it is most heavily used to form quantum wells in high speed transistors, and LEDs and lasers in the GaAs, InP, and GaN material families. In these structures, both electrons and holes would prefer to remain in the lowest bandgap region to lower their energy.

The **Type II**, or staggered alignment is typically seen in Si/Ge heterostructures. This alignment separates electrons and holes: for example in the Figure, holes would prefer to be in semiconductor A on the left, and electrons in semiconductor B on the right. To an extent, this structure mimics a p-n junction energy band diagram even when it is undoped.

The **Type III**, or broken gap alignment is seen for some combinations of semiconductors such as InAs/GaSb, etc. In such a heterojunction, the valence band electron wavefunctions on the left have an evanescent tail in the conduction band states on the right, and vice versa. A charge dipole is formed as a result naturally in such junctions.

The bottom two panels show how the energy bands of a heterojunction evolve with doping. The example chosen is a n-n heterojunction of Type I band offset with the respective Fermi levels fixed by doping on each side. Following the same strategy as the Schottky and pn junction diodes, we conclude that semiconductor B on the right will lose electrons to semiconductor A on the left. Semiconductor B will thus form a depletion region, and become positively charged due to the ionized donors left behind, and the energy band there will curve upwards as shown. However, semiconductor A was already n-type, and now has *more* electrons than was there due to doping. This sort of situation is called *accumulation* of carriers. Semiconductor A near the junction is negatively charged not by bound ionized acceptors (because there are none), but by the mobile electrons dumped into them from semiconductor B. Because of the negative charge, no matter what the origin, the band must curve downward as shown. It is also possible that the conduction band edge of semiconductor A goes below the Fermi level. In that case, semiconductor A can have 3D electrons far from the junction, but bound 2D states near the heterojunction. If semiconductor A was initially undoped, then the charge transfer will create a 2D electron gas at the interface. This is called *modulation doping*, in which carriers are "imported" from another region. We will discuss such doping and its uses in future chapters.

It is worth pointing out that since the exact nomenclature of naming band offsets as Types I, II or III (or straddling, staggered or broken) may sometimes vary in the literature based on the energy alignments

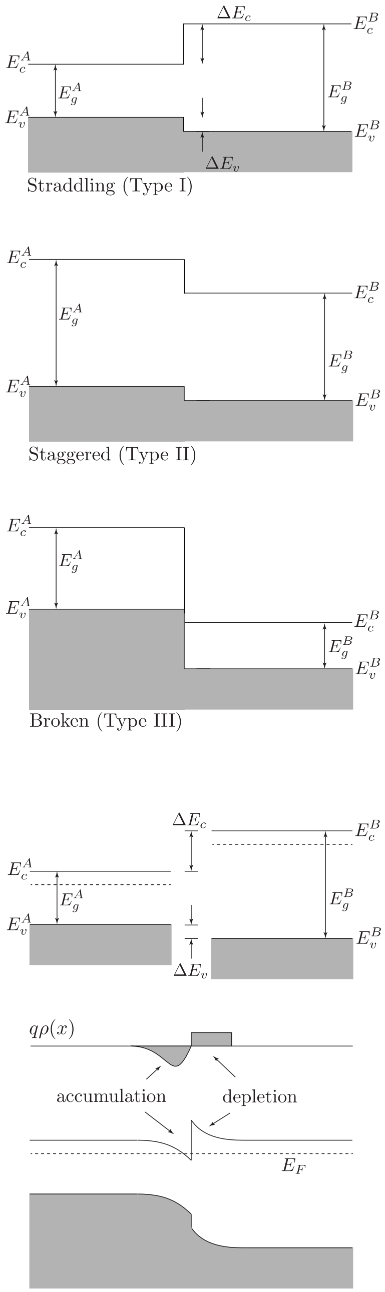
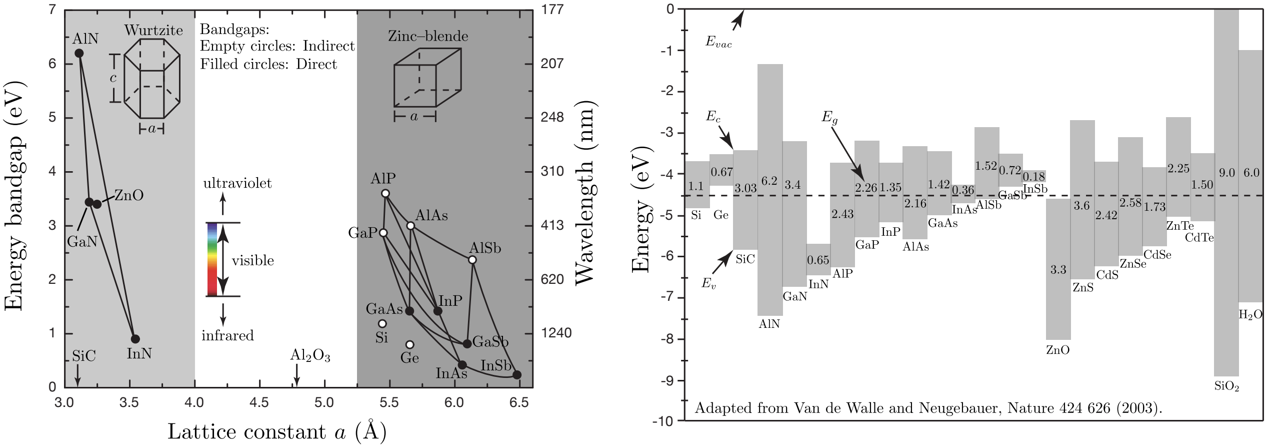


Fig. 15.13 Various types of band alignments (top three figures). Bottom two figures: an example of electrostatics and band-bending at an isotype n-n heterojunction.



Adapted from Van de Walle and Neugebauer, Nature 424 626 (2003).

Fig. 15.14 Bandgaps vs. lattice constants, and band alignments of 3D semiconductors.

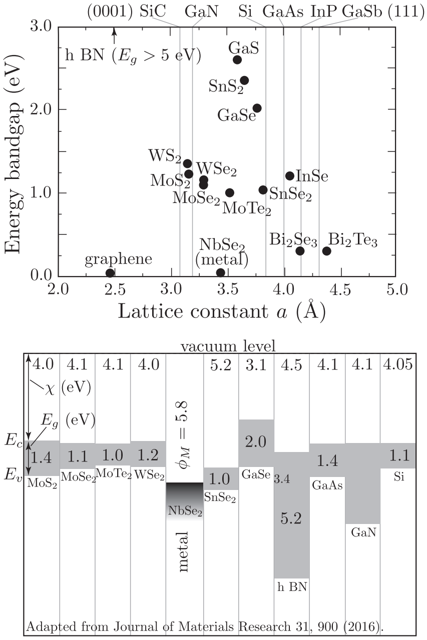
or sometimes the band symmetries, it must be clearly defined every time to avoid confusion.

15.6 Energy band diagrams: Poisson+Schrödinger

In this section, we show an example of the evaluation of the energy band-diagram of a realistic semiconductor heterostructure that serves to ilustrate all the concepts of this chapter. We will do this as a solved exercise. Consider the AlGaAs/GaAs quantum well semiconductor heterostructure shown in Fig. 15.16. Structures like this are used to make modulation doped field effect transistors (MODFETs), also called high electron-mobility transistors (HEMTs) that are used in cell phones and electronic communication systems. Except the gray region where the donor doping density is shown, all other regions are undoped, and the layer thicknesses are defined. The substrate is thick. The surface Schottky-barrier height is $q\phi_b = 0.6$ eV. All following questions are to be answered for a z-axis running vertically from the metal on the top into the substrate layer at the bottom. For semiconductor properties, energy bandgaps, and energy band offsets, a good source is the website `http://www.ioffe.ru/SVA/NSM/Semicond/`.

a) Using Gauss's law, calculate by hand the net mobile electron sheet density in the conduction band in the GaAs quantum well n_s in $10^{12}/\text{cm}^2$ units, the electric field in 10^6 V/cm units, and energy band diagrams in eV units by neglecting quantum effects completely. Sketch them as a function of the depth from the surface z in Å units.

We shall now compare this hand-calculation with a self-consistent Poisson–Schrödinger solution. For this problem, use a self-consistent Poisson–Schrödinger solver to calculate, plot, and answer the follow-



Adapted from Journal of Materials Research 31, 900 (2016).

Fig. 15.15 Bandgaps vs. lattice constants, and band alignments of 2D semiconductors compared to assorted 3D semiconductors.

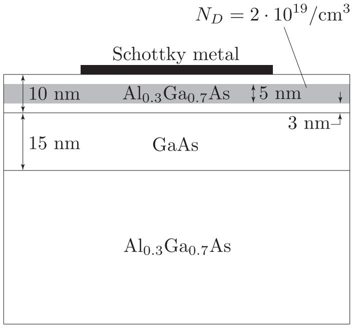

$N_D = 2 \cdot 10^{19}/\text{cm}^3$
Schottky metal
10 nm
$\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$
5 nm
3 nm
15 nm
GaAs
$\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$

Fig. 15.16 An AlGaAs/GaAs/AlGaAs quantum well heterostructure with a Schottky gate metal for Section 15.6.

ing:

b) The energy band diagram. Plot the conduction and valence band edges E_c, E_v and the Fermi level E_F in eV units as a function of the depth from the surface in Å units.

c) The mobile electron density $n(z)$ in the conduction band as a function of the depth from the surface in $10^{18}/\text{cm}^3$ units as a function of the depth from the surface in Å units.

d) The electric field $F(z)$ in 10^6 V/cm units as a function of the depth from the surface in Å units.

e) Plot the effective mass (or envelope) wavefunctions of the first five subbands in the GaAs quantum well as a function of the depth from the surface in Å units.

f) Explain the correlation between the electron density, the electric field, and the occupied energy subbands.

g) Compare the exact solution with the hand-calculation of part (a). Which quantities from the hand-calculation are accurate, and which ones are not trustworthy?

Solution:[14]

a) The dopants are in the $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ layer which is a wider-bandgap semiconductor than GaAs. As a result, the lowest energy states for the electrons are not at the location of the dopants, but in the GaAs layer, and in the Schottky metal on the surface. All the dopant electrons end up either in the QW, or the metal. We will use this for the hand calculation.

The total sheet density of dopants is $\sigma_d = N_D \cdot t_d = 10^{13}/\text{cm}^2$ where $t_d = 5$ nm is the thickness and $N_D = 2.10^{19}/\text{cm}^3$ is the doping density, which is positive in sign when ionized. Assuming the sheet density in the metal is σ_m and in the GaAs QW is σ_{qw}. Both are negative in sign, though σ_m and σ_{qw} are the magnitudes. Charge neutrality requires $\sigma_m + \sigma_{qw} = \sigma_d$, implying we need one more equation to find σ_{qw}.

This is found from the condition that the total voltage drop across the structure from the metal into the bulk of the semiconductor is equal to the built-in voltage. This voltage is the net area under the electric field that develops due to the transfer for charge between the metal and the semiconductor, given by the difference of work functions $\phi_s - \phi_m$.

Though this problem can be solved more accurately by hand, since we are doing a numerical evaluation in the next parts[15], let's do a lightning-fast estimation: assume all the dopants are scrunched up at the center of the doped layer a distance $t_1 = 4.5$ nm from the interface with the Schottky metal. Further, assume that all the 2DEG electrons are scrunched up $\sim t_3 \approx 5$ nm from the top $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$/GaAs interface $t_2 \sim 5.5$ nm from the dopant sheet. The 2DEG sheet is thus $t_2 + t_3 = 10.5$ nm from the dopant sheet. The electric field above the dopant sheet E_1 must point towards the metal, and below it E_2 towards the QW[16]. Using a built-in voltage of $\phi_b - \frac{1}{q}\Delta E_c$ and Gauss's law, we

---

[14] The basic method of Gauss's law in 1D problems states that the electric field due to an infinitely large area sheet charge of charge density σ is $E = \frac{\sigma}{2\epsilon_s}$ at all distances from it. For electric fields outside a sheet, one can use the integrated sheet charge of the volume. A corollary is that the electric field changes by a magnitude $E_2 - E_1 = \frac{\sigma}{\epsilon_s}$ across a sheet charge.

[15] A self-consistent 1D Poisson-Schrödinger solver is available from `https://www3.nd.edu/~gsnider/`.

[16] Electric fields always point from the positive to the negative charge.

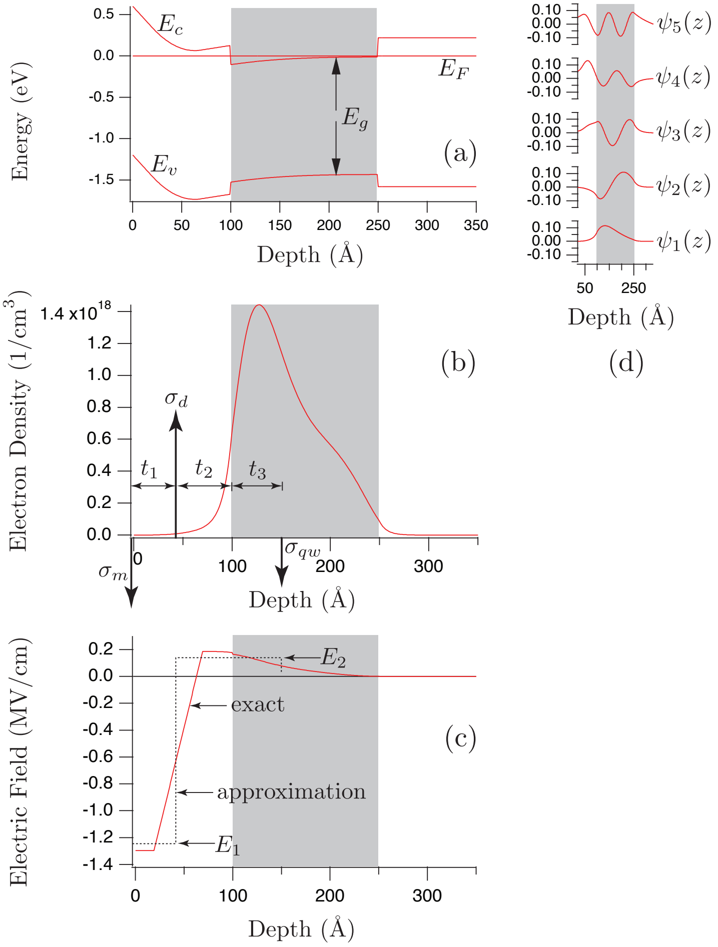


Fig. 15.17 Solution to Problem 15.6 produced using 1D Poisson simulator. Left, top: energy band diagram for part (b). Left, middle: mobile carrier density $n(z)$ for part (c). Left, bottom: Electric field $E(z)$ for part (d). Right: Effective mass-wavefunctions of the lowest 5 quantized states in the GaAs quantum well. Note that the actual electron-concentration $n(z)$ of Left, middle is similar in shape to the ground-state wavefunction $\psi_1(z)$, which is a strong indication that most of the mobile electrons in the quantum well are in the ground state. This is easily confirmed by finding that only one eigenvalue falls below the Fermi level, and all others are above it.

get

$$\begin{aligned}\sigma_m + \sigma_{qw}\sigma_d = N_D \cdot t_d \text{ [Charge Neutrality]} \\ -|E_1| \times t_1 + |E_2| \times (t_2 + t_3) = \phi_s - \phi_m \approx \phi_b - \Delta E_c \text{ [Total Potential Drop]} \\ E_1 = \frac{\sigma_m}{\epsilon_s} \text{ and } E_2 = \frac{\sigma_{qw}}{\epsilon_s} \text{ [Gauss's Law].}\end{aligned} \tag{15.26}$$

Solving the charge neutrality and potential drop equations immediately yields the electric fields and charges, whose signs are obtained by inspection:

$$\begin{aligned}
|E_1| = \frac{qN_d t_d}{\epsilon_s} \cdot \frac{t_2 + t_3}{t_1 + t_3 + t_3} + \frac{\phi_b - \Delta E_c}{t_1 + t_2 + t_3} &\implies \boxed{E_1 = -1.25 \text{ MV/cm}} \\
qn_m = \sigma_m = \epsilon_s E_1 &\implies n_m = \frac{\epsilon_s E_1}{q} \\
n_m = N_D t_d \left(\frac{t_2 + t_3}{t_1 + t_2 + t_3}\right) + \frac{\epsilon_s(\phi_b - \Delta E_c)}{q(t_1 + t_2 + t_3)} &\implies \boxed{n_m \approx 8.9 \times 10^{12}/\text{cm}^2} \\
|E_2| = \frac{qN_d t_d}{\epsilon_s} \cdot \frac{t_1}{t_1 + t_3 + t_3} - \frac{\phi_b - \Delta E_c}{t_1 + t_2 + t_3} &\implies \boxed{E_2 = +0.15 \text{ MV/cm}} \\
qn_{qw} = \sigma_s = \epsilon_s E_2 &\implies n_{qw} = \frac{\epsilon_s E_2}{q} \\
n_{qw} = N_D t_d \left(\frac{t_1}{t_1 + t_2 + t_3}\right) - \frac{\epsilon_s(\phi_b - \Delta E_c)}{q(t_1 + t_2 + t_3)} &\implies \boxed{n_{qw} \approx 1.1 \times 10^{12}/\text{cm}^2}
\end{aligned} \tag{15.27}$$

Compare the hand-calculated approximate fields and charges shown as a dashed line with the exact solutions as shown in Fig. 15.17: in spite of the very rough approximations, we are very close to the exact answers! Note that we could do much better even by hand, as we will see in later problems.

b) - e): Refer to Fig. 15.17.

f) The occupied electron subbands are those whose energy eigenvalues are below the Fermi level. Since only one eigenvalue falls below the Fermi level, that state is occupied with the highest density of electrons, and the net electron density shape also takes the form of the wavefunction squared of the ground state, as seen in Fig. 15.17, Left, middle and Right, bottom.

g) The electron sheet densities and electric fields can be obtained reasonably accurately by hand. This is because they primarily depend on Gauss's law and only weakly on the Schrödinger equation. If we want to know exactly how deep is the ground state in the quantum well or the shape of $n(z)$, the hand calculation will not be the most accurate in quantum structures, though even then a decent approximation can be obtained by approximating the shape of the quantum well. An infinite-well approximation is a good starting point for calculations by hand.

15.7 Polarization-induced doping in heterostructures

Semiconductors of the GaN, InN, and AlN family, and the ZnO and ZnMgO family have the hexagonal close packed (HCP) Bravais lattice. They have four atoms in the basis each (e.g. two Ga and two N for GaN), and are of the wurtzite crystal structure. Unlike the zinc-blende or diamond structure of most III-V and group IV semiconductors, the wurtzite structure is called uniaxial, because it has a unique (0001) c-axis along which it has a broken inversion symmetry. To see this, imagine physically flipping a GaN unit cell upside down along the c-axis: the atomic and bonding arrangement has changed. It is analogous to a bar magnet – the north and south poles of a magnet are tied to the ends, and flipping it upside down changes the magnetic field direction. GaN has a Ga-polar, and a N-polar direction: the (0001) direction is Ga-polar, and the $(000\bar{1})$ is the N-polar. This is not the case in GaAs or Si. Now because of the large electronegativity of the chemical bonds, the bonds are partially ionic, and the lack of a center of symmetry results in the formation of an internal electric field in the crystal, *even without any dopants*! This is called spontaneous polarization. The quantum mechanical explanation of this phenomena uses the Berry phase discussed in Chapter 9.

Here, we discuss briefly the consequences of the presence of polarization in semiconductors. Fig. 15.18 shows that if the microscopic charge dipoles $\mathbf{p}_i$ formed in partially ionic sp^3 tetrahedral bonds do not exactly cancel, we have in every tetrahedron a charge dipole left over, resulting in a net macroscopic polarization $\mathbf{P} = \sum_i \mathbf{p}_i/\Omega$, where Ω is the volume of the unit cell. In a GaN crystal, the tips and the tails of these dipoles are in the c-planes – though in the bulk these charges balance and cancel, this is impossible on the surface. At the surface the entire polarization charge appears as a *bound sheet charge* of density $\sigma_\pi = \mathbf{P} \cdot \hat{\mathbf{n}}$. On the top Ga-face of the crystal, the polarization sheet charge is $-\sigma_\pi$, and an exactly equal and opposite charge density $+\sigma_\pi$ must appear at the other surface of the crystal, which is the N-face.

Because of the appearance of these bound charges, there is an electric field $F_\pi = \sigma_\pi/\epsilon_s$ set up in the bulk, which causes the bands to bend, as seen in the figure. At heterojunctions, the charge density that appears at the interface is the *difference* of the polarization fields across: $\sigma_\pi = (\mathbf{P}_1 - \mathbf{P}_2) \cdot \hat{\mathbf{n}}$. As a result of these interface polarization charges, the bands can bend in a way that mobile 2D electron or hole gases are *induced*. Since the valence band is a reservoir of electrons, the conduction band at the heterointerface can be filled with electrons pulled out of the valence band near the surface due to the polarization field. Similarly, the electrons from the valence band can be pushed out to create holes in the bulk. If these junctions are close to the surface, filled and empty surface states also participate in this form of polarization-induced doping. Since this form of doping is created by electrostatics rather than thermal ionization, the mobile carrier densities generated are nearly temperature independent.

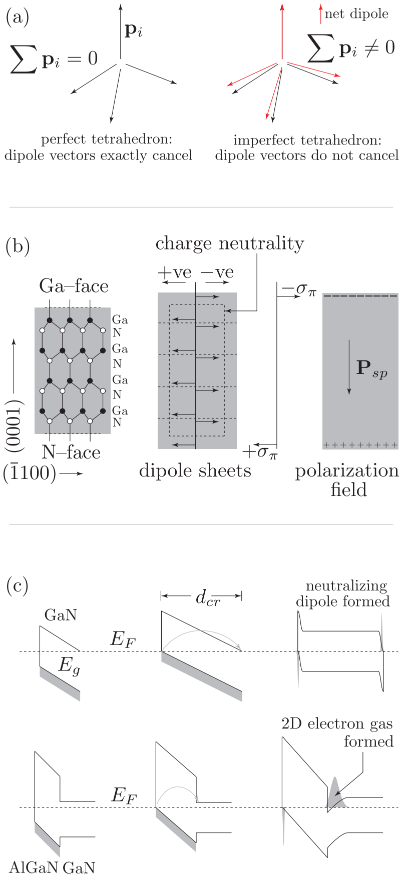

(a)
$\sum \mathbf{p}_i = 0$
$\mathbf{p}_i$
net dipole
$\sum \mathbf{p}_i \neq 0$
perfect tetrahedron: dipole vectors exactly cancel
imperfect tetrahedron: dipole vectors do not cancel
(b)
charge neutrality
+ve −ve
Ga–face
Ga N Ga N Ga N Ga N
(0001)
N–face
$(\bar{1}100)$
dipole sheets
$-\sigma_\pi$
$+\sigma_\pi$
$\mathbf{P}_{sp}$
polarization field
(c)
GaN
E_F
E_g
d_{cr}
neutralizing dipole formed
2D electron gas formed
E_F
AlGaN GaN

Fig. 15.18 Energy band diagrams of polar heterostructures such as GaN/AlGaN heterjunctions. because of the presence of internal electric charges due to spontaneous and piezoelectric polarization, electric fields and band bending happens *even in the absence* of donor or acceptor doping.

15.8 Chapter summary section

In this chapter, we learned:

- The carrier statistics of an intrinsic semiconductor, and the mass action law $np = n_i^2$ at thermal equilibrium.
- How to relate the bandstructure of a semiconductor to its carrier statistics: for example, how to handle multiple conduction band valleys in determining the mobile electron densities.
- How charge neutrality defines the location of the Fermi level in all semiconductors.
- How donor and acceptor doping converts a semiconductor into an n-type or p-type conductor, and how to properly handle the many-particle aspect of donor and acceptor statistics.
- How to draw energy band diagrams of metal-semiconductor Schottky junctions, pn junctions, and heterojunctions using the universal philosophy that the Fermi level is the same everywhere. The energy band diagram is drawn by tracking the flow of charge $\rightarrow$ find the electric field $\rightarrow$ draw the energy band diagram.
- A rich range of energy band alignments for heterostructures of 3D and 2D semiconductors.
- How electronic polarization can create unbalanced charges and fields without doping.

Further reading

Most books on semiconductor *devices* cover the topics covered in this chapter in sufficient detail, because of the central importance of the control of electrons and holes in device applications. The popular book **Physics of Semiconductor Devices** by Sze and Ng is a good reference, as are **Device Electronics for Integrated Electronic Circuits** by Muller, Kamins, and Chan and **Fundamentals of Modern VLSI Devices** by Taur and Ning. The old and timeless classics are Shockley's **Electrons and Holes in Semiconductors**, Spenke's **Electronic Semiconductors**, and **Physics and Technology of Semiconductor Devices** by Grove (the founder of Intel!) – but good luck getting these books as they are chronically out of print!

Exercises

(15.1) **The deep-acceptor problem and the 2014 Physics Nobel Prize**

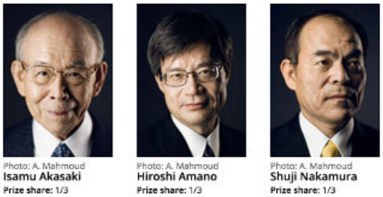


Fig. 15.19 The 2014 Physics Nobel Prize went to the researchers who solved the p-type doping problem in GaN. This work made quantum well blue LEDs and lasers, and LED lighting possible.

(a) Show that for a homogeneous semiconductor with electron mobility μ_n and hole mobility μ_p, the *lowest* conductivity that can be achieved at thermal equilibrium is $\sigma_{min} = 2qn_i\sqrt{\mu_n\mu_p}$, irrespective of the donor or acceptor doping.

Magnesium is a relatively deep acceptor in the wide bandgap semiconductor GaN. The acceptor ionization energy is $E_A \sim 160$ meV. Consider a GaN sample ($E_g = 3.4$ eV, $m_c^\star \sim 0.2m_0$, $m_v^\star \sim 1.4m_0$) doped with $N_A = 10^{18}/\text{cm}^3$ magnesium acceptors. In the process of doping this sample with magnesium, unintentional donors of density $N_D = 10^{14}/\text{cm}^3$ of donor ionization energy $E_D = 10$ meV also incorporate into the semiconductor.

(b) For T=300 K, plot the log of n, p, N_A^-, N_D^+, $n + N_A^-$, and $p + N_D^+$ as a function of the Fermi level E_F. Remember the Fermi level can be within the gap, or in the conduction or valence bands. So in your plot vary the values of E_F from below E_v to above E_c. Indicate the donor and acceptor ionization energies and show in the plot where the real Fermi level at 300 K is. Explain.

(c) What are the densities and types of mobile carriers in the sample at 300 K? Is the sample n- or p-type? Find the conductivity of the sample at 300 K if the electron mobility is $\mu_n \sim 1000\ \text{cm}^2/\text{V·s}$ and the hole mobility is $\mu_p \sim 10\ \text{cm}^2/\text{V·s}$.

(d) Do online research of the connection between the p-type doping problem of wide-bandgap semiconductors and the 2014 Physics Nobel Prize and write a short summary of what you find.

(15.2) **Doping a semiconductor above the Mott criteria** In this problem, we will find the critical doping density of a semiconductor above which it's electrical conductivity starts behaving like a metal. The signature of metallic conduction is when the mobile electron density becomes nearly independent of temperature. Consider GaAs, a 3D semiconductor crystal with a bandgap of $E_g = 1.4$ eV, a dielectric constant of $\epsilon_s = 13\epsilon_0$ and a conduction band-edge effective mass of $m_c^\star = 0.067m_e$. We dope the semiconductor with donor atoms to a 3D density N_d, and investigate the free electron density in the conduction band.

(a) Which dopants atom(s) from the periodic table will you choose to make GaAs n-type? Which site of the crystal should the atom(s) of your choice sit to do the job?

(b) Find the donor ionization energy based on the effective mass theory.

(c) Assume a doping density of roughly $N_{d1} = 10^{17}/\text{cm}^3$. Argue why at room temperature $T = 300$ K, the mobile electron density in the conduction band may be approximated to be roughly equal to the doping density $n \approx N_{d1}$. Under this assumption, find the Fermi level E_F with respect to the conduction band edge E_c for the doping density N_{d1}.

(d) The mobile electron density is measured as a function of temperature from $T \sim 300$ K to $T \sim 1$ K. *Qualitatively* sketch how the electron density will change with the lowering of temperature. Also *qualitatively* sketch how will the electrical conductivity of the doped GaAs change in the same temperature range by including a rough dependence of electron mobility.

(e) Find the effective Bohr radius $a_B^\star$ of a donor electron state using the effective mass theory. This is a characteristic length scale over which the wavefunction of the electron due to the donor spreads in real space, with the center at the donor site. You should get an answer roughly equal to $a_B^\star \approx 10$ nm.

(f) Now consider the Mott critical doping density $N_d^{cr} = \frac{1}{(a_B^\star)^3}$. Explain why at this, and higher doping densities, it is not necessary for the electron to be thermally ionized to the conduction band to become mobile in the crystal. How would the electron move in the crystal under this situation?

(g) The above critical doping density is called the Mott criterion, up to a constant of order unity. If the doping density in GaAs is $N_{d2} > N_d^{cr}$, re-do your answer for part (d) by sketching qualitatively the dependence of the mobile electron density, and the electrical conductivity as a function of temperature. Explain why this is a characteristic of a metal, and heavy doping has effectively converted an insulator (or semiconductor) to a metal.

(h) Based on the above, explain why narrow bandgap semiconductors are prone to become metallic at low levels of unintended donor impurities, meaning to ensure they have semiconducting properties, they must be much purer than their wider-bandgap counterparts.

(15.3) **The mystery of high p-type conductivity in AlN**

Here is a mystery you can help solve. Now aluminum nitride (AlN) is one of the widest bandgap semiconductors we know, with an energy gap of nearly 6.2 eV. The bandstructure is shown in Fig. 13.12 of the text. It is notoriously difficult to dope AlN n-type, and most people today consider impossible to dope p-type. If you can dope it p-type, it will be a major breakthrough both for photonic and electronic devices.

In 2014 a graduate student in my group was investigating thin layers of AlN deposited on top of an insulating silicon crystal by Hall-effect measurements of the conductivity. He called my cellphone excitedly late one night saying that he had measured very high p-type hole concentration and hole mobility in the sample, and therefore very high p-type conductivity. Turns out he was correct, but alas, it was not exactly the breakthrough of AlN p-type doping we were dreaming about. What might have happened that gave such a high p-type conductivity for these AlN on silicon samples?

(15.4) **Semiconductor heterojunctions and band offsets: know what you are talking about**

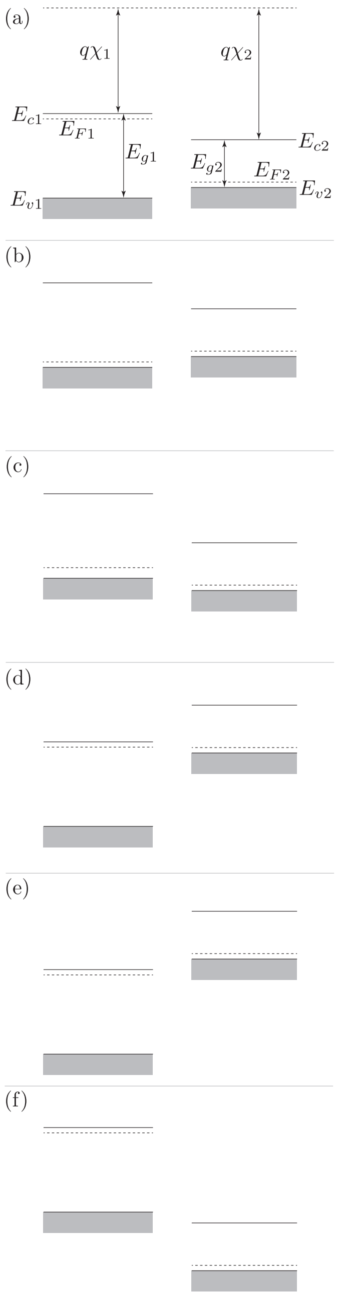


Fig. 15.20 Various heterojunctions before reaching equilibrium.

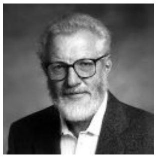


Fig. 15.21 Herbert Kroemer, a pioneer in semiconductor heterojunction based design of transistors and lasers was awarded the 2000 Nobel Prize in Physics.

Here is Kroemers [see Fig 15.21] Lemma of Proven Ignorance: *If, in discussing a semiconductor problem, you cannot draw an* **energy band diagram**, *this shows that you don't know what you are talking about*, with the corollary: *If you can draw one, but don't, then your audience won't know what you are talking about.*

In this problem we make sure we don't fall into this trap!

In Fig. 15.20 (a-f), energy band diagrams of several different semiconductor heterojunctions are shown (one on the left side and the other on the right side) *before* the two materials are put together, with their individual Fermi levels indicated. In each case sketch the equilibrium energy band diagram when a heterojunction is formed between the two semiconductors. In each case indicate the depletion and/or accumulation and/or inversion regions that may exist in equilibrium on either side of the heterointerface. The alignment shown corresponds to the electron affinity rule as shown explicitly in part (a). All the labels are also shown in more detail in part (a), and which you can use for the other parts.

(15.5) **High electron mobility transistors (HEMTs)**
Consider the AlGaAs/GaAs HEMT structure in Fig. 15.22. The structure is grown by MBE. The thicknesses and doping of the layers are: the cap layer: $t_{cap} = 5$ nm and $N_D = 7 \times 10^{17}/\text{cm}^3$, $\text{Al}_{0.3}\text{Ga}_{0.7}\text{As}$ layer thickness $t_1 = 25$ nm, after the gate-recess etch, AlGaAs thickness $t_2 = 17$ nm, δ-doped layer of thickness = 1 nm and effective 3D doping $= 3.5 \times 10^{19}/\text{cm}^3$, $t_3 = 5$ nm, and the GaAs quantum well thickness $t_4 = 10$ nm. Assume the surface barrier height is pinned at $q\Phi_s = 0.6$ eV below the conduction band edge for both GaAs and AlGaAs.

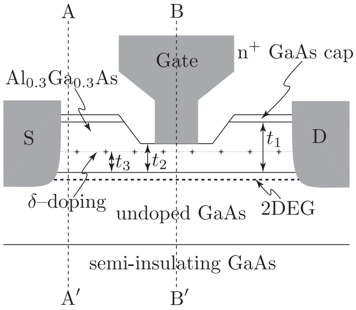


Fig. 15.22 AlGaAs/GaAs δ-doped HEMT.

(a) Calculate the 2DEG sheet density in the GaAs QW below the gate. Draw the charge-field-band diagram along line B-B′ for finding the sheet density. Verify your calculated value with 1D Poisson–Schrödinger simulation of the charge-field-band diagram. Is there any quantum-confinement? How many quantum-confined states are formed in the GaAs QW? What are the eigenvalues?.

(b) Calculate the 2DEG sheet density in the GaAs QW below the source-and drain-access regions. Draw the charge-field-band diagram along line A-A′ for finding the sheet density. Verify your calculated value with 1D Poisson simulation of the charge-field-band diagram. Comment on quantum confinement and eigenvalues.

(c) What is the gate-capacitance C_g? Verify your analytical result with a 1D Poisson–Schrödinger simulation of C_g vs. the applied gate voltage V_g.

(d) Calculate the threshold voltage V_{th} for the HEMT. Verify your analytical result with a 1D Poisson–Schrödinger simulation of the sheet density n_s vs. the applied gate voltage V_g; plot this in both linear and log scales.

(15.6) **Energy band diagrams of a heterostructure bipolar transistor (HBT)**
Fig. 15.23 shows the energy band alignments of three semiconductors with respective dopings that will be put together in the same order as the figure into a heterostructure bipolar transistor (HBT). Draw the energy band diagram of the HBT at equilibrium.

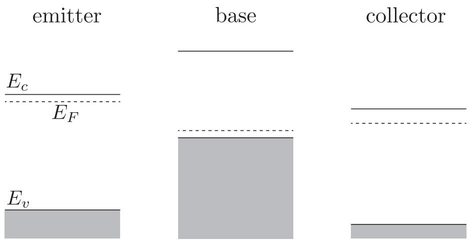


Fig. 15.23 Band alignments in a heterostructure bipolar transistor.

(15.7) **Polar III-nitride semiconductor heterostructures**

Consider an $\mathrm{Al_{0.3}Ga_{0.7}N/GaN}$ HEMT structure. Assume that the surface Schottky barrier height is $q\phi_s = 1.7$ eV on AlGaN and $q\phi_s = 0.9$ eV on GaN. The $\mathrm{Al_{0.3}Ga_{0.7}N}$ layer is coherently strained to the GaN lattice constant.

(a) Calculate the net polarization sheet charge $Q_\pi(x)$ at a strained $\mathrm{Al_xGa_{1-x}N/GaN}$ heterojunction for barrier Al composition x. Use the value for $x = 0.3$ for the rest of the problem.

(b) How does the *mobile* sheet charge at the AlGaN/GaN junction vary with the thickness t_b of the AlGaN barrier? Plot the sheet charge n_s for AlGaN thicknesses up to $t_b = 40$ nm.

(c) Plot the energy band diagram of an AlGaN/GaN HEMT with a $t_b = 30$ nm AlGaN cap at zero gate bias, and at pinch-off. What is the pinch-off voltage? Verify your analytical calculation with a self-consistent 1D Poisson–Schrödinger solution.

(d) Now, a $t_{cap} = 5$ nm layer of GaN is added above the AlGaN barrier. Calculate and plot the band diagram of this structure at zero bias and at pinch-off. What is the effective Schottky-barrier height in these two cases? Do you expect the gate leakage of this diode to be different from the AlGaN/GaN structure? Why (not)?

(15.8) **Thermodynamics: the great leveler**

In semiconductor device physics, the central tenet is that the Fermi level is the same everywhere in thermal equilibrium. We proved it in statement 15.17 of this chapter. Think deeper about what this means – because it is tied to the concept of thermal equilibrium itself.

(a) The statement does not require that the flow of electrons across the junction be zero in any direction – just that it be *equal in both directions* such that the *net flow is zero*. This implies that based on the DOS of each side of the junction, there can be currents flowing in both directions. Sketch these currents for a few junctions and examine their dependence on the Fermi level, temperatures, and band alignments.

(b) A more interesting question is: if the junction was out of equilibrium due to an applied voltage or optical illumination, and the perturbation is suddenly turned off, how long does it take for it to reach thermal equilibrium? Argue why the currents you have sketched above are responsible to return the system to thermal equilibrium, and their microscopic nature will determine the duration of the return.

The discussion above highlights the following important fact. *The final equilibrium state itself is independent of the nature of the microscopic details that take it to that state. But how the system gets into, and out of, the equilibrium state, is very much dependent on the microscopic details!*

Controlling Electron Traffic in the k-Space

16

This chapter marks a summary of Modules I and II, and the beginning of a new adventure: Modules III and IV, on quantum electronics and quantum photonics with semiconductors. The goal of this chapter is to serve as a roadmap by making a smooth transition from Modules I and II. This is achieved by first *succinctly* summarizing the physics of semiconductors developed in the first two modules. Then, the quantum electronics and photonics of Modules III and IV are also succinctly summarized. There are significant advantages to sampling the flavor of all the core concepts of the subject in one place. You will solidify the concepts you have learned in Modules I and II, and should come back and re-read this chapter after going through Modules III and IV to see the big picture.

- The first part of this chapter succinctly discusses the quantum physics of electron and hole statistics in the bands of semiconductors, the quantum mechanical transport of the electron and hole states in the bands, and optical transitions between bands. The unique point of view presented here is a unified picture exemplified by single expressions for the carrier statistics, transport, and optical transitions for electrons and holes in nanostructures of all dimensions – ranging from bulk 3D, to 2D quantum wells, to 1D quantum wires, and with a direct connection to electronic and photonic devices.
- The second part of the chapter summarizes the core concepts of the physics, and electronic and photonic applications of semiconductors and nanostructure based devices such as diodes, transistors, solar cells, light emitting diodes, and lasers.
- Drawing an analogy, Modules I and II tell us how to create the *architecture* of a city – but for electrons: with houses and parking lots carved out of bands and heterostructures, roads of electron bandstructure $E(k)$ and speed limits of $v_g(k) = \frac{1}{\hbar}\nabla_k E(k)$ for each k-lane. In Modules III and IV we learn the tools of traffic control: gates and one-way lanes, electron dynamics in response to voltages and light, and how to coax out useful semiconductor devices by combining these conceptual building blocks.

16.1 Electron energies in semiconductors 338

16.2 Semiconductor statistics 340

16.3 Ballistic transport in semiconductors 342

16.4 Ballistic transport in non-uniform potentials/tunneling 345

16.5 Scattering of electrons by phonons, defects and photons 347

16.6 The Boltzmann transport equation 351

16.7 Current flow with scattering: drift and diffusion 354

16.8 Explicit calculations of scattering rates and mobility 358

16.9 Semiconductor electron energies for photonics 361

16.10 The optical joint density of states $\rho_J(\nu)$ 363

16.11 Occupation of electron states for photonics 366

16.12 Absorption, and emission: spontaneous and stimulated 367

16.13 Chapter summary section 371

Further reading 371

Exercises 371

16.1 Electron energies in semiconductors

As we have discussed in Chapters 1–15, electrons in free space are allowed to have *continuous* values of energies $E(k) = \frac{\hbar^2 k^2}{2m_e}$, where $\hbar = h/2\pi$ is the reduced Planck's constant, m_e the rest mass of an electron, and $k = 2\pi/\lambda$ is the wavevector with λ the electron wavelength. $\hbar k = \frac{h}{\lambda} = p$ is the momentum of the free–electron by the de Broglie relation of wave-particle duality. The dynamical properties of the electron are governed neither by Newton's law of particles, nor Maxwell's equations for waves, but by the Schrödinger equation, which has the wave-particle duality baked in from the get-go.

A periodic potential $V(x + a) = V(x)$ in the real space of a crystal of lattice constant a, when included in the Schrödinger equation, is found to split the continuous energy spectrum electron energies $E(k) = \frac{\hbar^2 k^2}{2m_e}$ into bands of energies $E_m(k)$, separated by energy gaps, shown in Fig. 16.1. The m^{th} allowed energy band is labeled $E_m(k)$. The states of definite energy $E_m(k)$ have real-space wavefunctions given by $\psi_k(x) = e^{ikx} u_k(x)$ called Bloch functions, where $u_k(x + a) = u_k(x)$ is periodic in the lattice constant. Each band has exactly the same number of allowed electron states N as the number of unit cells in the real space[1]. Each state of energy $E_m(k_n)$ in each band m is indexed by a unique $k_n = \frac{2\pi}{L} n$ with $n = 0, \pm 1, \pm 2$, where $L = Na$ is the macroscopic size of the crystal. The k_n lies in the first Brillouin zone, between $-G/2 \leq k \leq +G/2$, where $G = 2\pi/a$ is the reciprocal lattice vector. The Pauli exclusion principle allows each state to be occupied by $g_s = 2$ electrons of opposite spins. Thus, each band $E_m(k)$ can hold a maximum of $2N$ electrons. The energy bands have the property $E_m(k + G) = E_m(k)$, i.e., they are periodic in the k-space.

The k of the band $E_m(k)$ has a different meaning from the wavevector of a free–electron. $\hbar k$ is not the momentum of the electron of a unique energy $E_m(k)$. States of definite energy (or energy eigenstates) $E_m(k)$ are *not* states of definite momentum in a periodic crystal[2]. The Bloch states are a mixture of states of momenta $\hbar(k \pm G)$, which is why $\hbar k$ is referred to as the "crystal momentum".

The group velocity of state k in energy band $E(k)$ is $\mathbf{v}_g = \frac{1}{\hbar} \nabla_{\mathbf{k}} E(\mathbf{k})$. In response to an external force $\mathbf{F}$, the crystal momentum changes according to $\mathbf{F} = \hbar \frac{d\mathbf{k}}{dt}$; the energies and velocities change to satisfy this relation.

The two statements above are the most remarkable results of the quantum mechanics of electrons in crystals, and are by no means obvious. The states at the very bottom *and the very top* of any band $E_m(k)$ must then have zero group velocity. An empty band cannot carry a charge current because there are no electrons in them. A filled band *also* cannot carry a net charge current – but for a different reason. In a filled band, the current carried by a filled state k has an exactly equal partner, but opposite in sign at $-k$. The net current cancels and gives zero: $J_{filled} = \frac{(-q)}{L^d} \sum_k f(k) v_g(k) = \frac{(-q)}{L^d} \sum_k v_g(k) = 0$ for a filled band,

[1] This is expected since if the atoms were very far apart before forming a crystal, the number of states at each allowed energy is simply equal to the number of atoms.

[2] A crude classical analogy is a ball moving at a constant velocity in a flat plane will maintain its momentum. But if it rolls in a periodic potential, its velocity increases and decreases periodically: its momentum is not fixed.

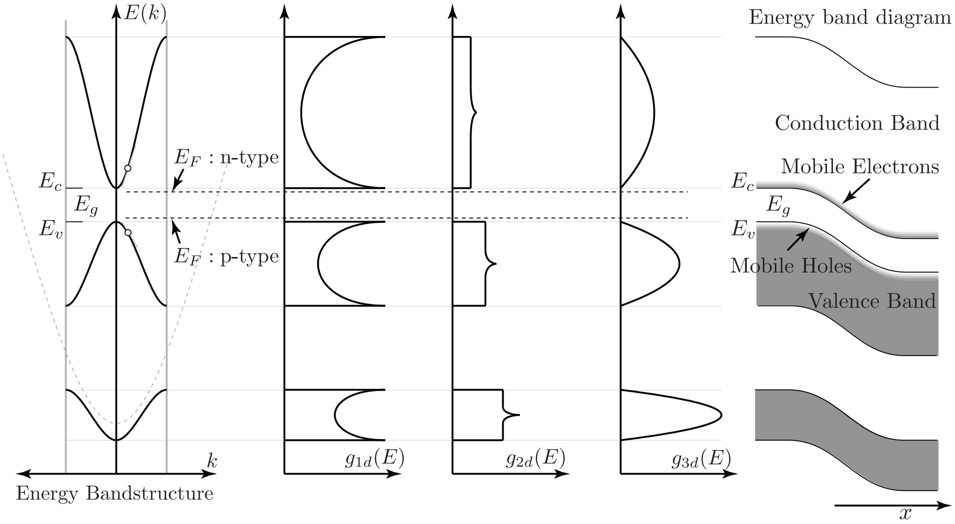


Fig. 16.1 Allowed electron energies in a crystal is the bandstructure $E(k)$. The DOS depends on the dimensionality: the 1D, 2D, and 3D DOS are shown. The Fermi level E_F is controlled by doping, electrostatically, or with light or heat. The band edges may be changed in real space also by doping, electrostatically, light or heat, leading to energy band diagrams.

where $f(k) = 1$ is the occupation probability of state k, $-q$ is the electron charge, and L^d is the volume in d dimensions. An empty band can carry current only if electrons are put into it – by doping, electrostatically, optically, or thermally. A filled band can carry a net current only if electrons are *removed* from it – again, by doping, electrostatically, optically, or thermally. When electrons are removed from the top of a filled band, the net current is

$$J = \frac{(-q)}{L^d} \sum f(k) v_g(k) = \frac{(-q)}{L^d} \sum 1 \cdot v_g(k) + \frac{(-q)}{L^d} \sum_k [f(k) - 1] \cdot v_g(k)$$
$$= \frac{(+q)}{L^d} \sum_k [1 - f(k)] v_g(k), \quad (16.1)$$

implying the current is effectively carried by "holes" or empty electron states, which behave in transport and electrostatically as positive mobile charges $+q$, where $-q$ is the normal electron charge. Fig. 16.1 shows mobile electrons at the bottom of the conduction band, and mobile holes as empty states at the top of the valence band.

A crystal is an **intrinsic semiconductor** if there is a few eV bandgap between a completely filled valence band $E_v(k)$ and the next conduction band $E_c(k)$ that is completely empty. At the very bottom of the conduction band, we can expand the energies as the parabola $E_c(k) \approx E_c + \frac{\hbar^2 k^2}{2m_c^\star}$, where $E_c(0) = E_c$ is the band-edge, or lowest energy, and $m_c^\star$ is the conduction band edge effective mass. Similarly, for the top of the valence band is approximated as $E_v(k) = E_v - \frac{\hbar^2 k^2}{2m_v^\star}$, with $E_v(0) = E_v$ the valence band edge, and $m_v^\star$ the valence band effective mass. $E_c - E_v = E_g$ is the energy bandgap of the semiconductor. If the crystal momentum k at which the conduction band reaches the minimum and the valence band reaches its maximum are exactly the same, we have a **direct bandgap** semiconductor, and if they are not, it is an **indirect bandgap** semiconductor.

16.2 Semiconductor statistics

We will build the **map of semiconductor physics** shown in Fig. A.1 in the Appendix. Some general results for semiconductor structures are now discussed, in which electrons and holes are allowed to move in d dimensions. $d = 3$ is the case in an ordinary 3D bulk semiconductor, $d = 2$ is in 2D semiconductor membranes or heterostructure quantum wells, $d = 1$ is in quantum wires, and $d = 0$ i sin quantum dots. The density of states near the edge of the parabolic conduction band in d-dimensions is given by

$$g_c^d(E) = \frac{g_s g_v}{(4\pi)^{\frac{d}{2}} \Gamma(\frac{d}{2})} (\frac{2m_c^\star}{\hbar^2})^{\frac{d}{2}} (E - E_c)^{\frac{d-2}{2}}, \tag{16.2}$$

where $g_s = 2$ is the spin-degeneracy for conventional situations, and g_v is the valley degeneracy. The units are in $1/(\text{eV} \cdot \text{cm}^d)$. The valley degeneracy of the conduction band is $g_v = 1$ for most direct-bandgap semiconductors in 3D, but is $g_v = 6$ for 3D silicon, and $g_v = 4$ for 3D Germanium, and $g_v = 2$ in single layer 2D graphene, BN, or transition metal dichalcogenide semiconductors. $\Gamma(...)$ is the gamma function, with $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$, and $\Gamma(n+1) = n\Gamma(n)$. For $d = 3$ the DOS increases with energy E as $\sqrt{E - E_c}$, for $d = 2$, the DOS is constant, and for $d = 1$, the DOS decreases as $1/\sqrt{E - E_c}$. The dependence on the conduction band edge effective mass is $(m_c^\star)^{\frac{d}{2}}$, which means a conduction band with a heavier effective mass has a higher DOS in *all* dimensions. Exactly in the same way, the valence band DOS has $m_v^\star$ for the mass, and the energy dependence on dimensions is $(E_v - E)^{\frac{d-2}{2}}$, the same as for the conduction band, except the argument is $E_v - E$ for obvious reasons at the top of the valence band. Fig. 16.1 shows the DOS of the *entire* bands, the approximations of Equation 16.2 are applicable near the band edges, where most carrier transport occurs.

At equilibrium, the Fermi level E_F determines the number of electrons in the conduction band n_d and the number of holes p_d in the valence band. On the other hand, at equilibrium, if we know either the number of electrons in the conduction band n_d or holes p_d in the valence band, the Fermi level E_F is uniquely determined. There is a one-to-one correspondence between the densities and E_F at equilibrium. The relation is obtained by using the Fermi–Dirac occupation function with the density of states:

$$n_d = \int_{E_c}^{\infty} dE \cdot g_c^d(E) \cdot \frac{1}{1 + e^{\frac{E_c(k) - E_F}{k_b T}}} = N_c^d \cdot F_{\frac{d-2}{2}}(\frac{E_F - E_c}{k_b T}). \tag{16.3}$$

The prefactor $N_c^d = g_s g_v (\frac{2\pi m_c^\star k_b T}{\hbar^2})^{\frac{d}{2}}$ is called the effective band-edge DOS, and has units of $1/\text{cm}^d$, where d is the dimension. For typical semiconductors, $N_c^{1d} \sim 10^6/\text{cm}$, $N_c^{2d} \sim 10^{12}/\text{cm}^2$, and $N_c^{1d} \sim 10^{18}/\text{cm}^3$ at $T = 300$ K. The dimensionless Fermi–Dirac integral of order j is $F_j(\eta) = \frac{1}{\Gamma(j+1)} \int_0^\infty du \frac{u^j}{1+e^{u-\eta}}$, the values of which may be obtained from tables, by calling functions (in Matlab, Mathematica, Python, or other packages), or by direct integration. Two important limiting cases of the Fermi–Dirac integral are the (semi-classical) **non-degenerate** Maxwell–Boltzmann limit $F_j(\eta) \approx e^\eta$ for $\eta << -1$, and the (quantum) **degenerate** limit $F_j(\eta) \approx \frac{\eta^{j+1}}{\Gamma(j+2)}$ for $\eta >> +1$. The dimensionless argument is $\eta = \frac{E_F - E_c}{k_b T}$, which is a measure of how far the Fermi-level E_F is from the band edge E_c, measured in units of the thermal energy $k_b T$. If for example, the Fermi level is at $E_F = E_c - 6k_b T$ for a 3D semiconductor with $N_c^{3d} = 10^{18}/\text{cm}^3$, $\eta = \frac{E_F - E_c}{k_b T} = -6$, and the electron density in the conduction band is $n_{3d} = N_c^{3d} \cdot F_{\frac{1}{2}}(\eta) \approx N_c^{3d} \cdot e^{-6} \approx \frac{10^{18}}{400} \approx 2.5 \times 10^{15}/\text{cm}^3$. On the other hand, if for the same semiconductor $E_F = E_c + 3k_b T$, $\eta = +3$, and the electron density is $n_{3d} = N_c^{3d} \cdot F_{\frac{1}{2}}(\eta) \approx N_c^{3d} \cdot \frac{\eta^{\frac{1}{2}+1}}{\Gamma(\frac{1}{2}+2)} = 10^{18} \cdot \frac{3^{\frac{3}{2}}}{\frac{3\sqrt{\pi}}{4}} \sim 4 \times 10^{18}/\text{cm}^3$. Similar arguments carry over to find the statistics and densities of holes in the valence band.

In the absence of dopants and impurities, an intrinsic semiconductor is **charge neutral**. This implies if there are any electrons thermally or optically ionized from the valence to the conduction band[3], it must leave behind an exactly equal number of holes in the valence band. This sets $n_d = p_d$, or $N_c^d F_{\frac{d-2}{2}}(\frac{E_F - E_c}{k_b T}) = N_v^d F_{\frac{d-2}{2}}(\frac{E_v - E_F}{k_b T})$, which is only possible for an unique E_F. **The charge neutrality condition fixes the location of the Fermi level** E_F. The product of the intrinsic carrier densities is

$$n_d \cdot p_d = N_c^d \cdot N_v^d \cdot F_{\frac{d-2}{2}}(\frac{E_F - E_c}{k_b T}) \cdot F_{\frac{d-2}{2}}(\frac{E_v - E_F}{k_b T}) \approx N_c^d \cdot N_v^d \cdot e^{-\frac{E_g}{k_b T}}, \tag{16.4}$$

where the approximation on the right holds only for non-degenerate carrier distributions, when $\frac{E_F - E_c}{k_b T} << -1$, *and* $\frac{E_v - E_F}{k_b T} << -1$. This condition is only met if E_F is far inside the gap, several $k_b T$ far from the band edges. In this case, the intrinsic electron or hole density in d-dimensions is obtained explicitly as $n_{di} = p_{di} = \sqrt{N_c^d N_v^d} \cdot e^{-\frac{E_g}{2k_b T}}$, and is exponentially dependent on the bandgap. For example, in 3D silicon of $E_g = 1.1$ eV at $T = 300$ K, $n_i \approx 10^{10}/\text{cm}^3$, whereas for GaN with $E_g = 3.4$ eV the intrinsic density is a measly[4] $n_i \approx 10^{-10}/\text{cm}^3$!

In the presence of dopants and impurities, at equilibrium a bulk semiconductor is *still* charge neutral. For a donor density N_D and donor state at energy E_D (meaning the ionization energy is $E_g - E_D$, using the valence band edge as the reference), and an acceptor density N_A and acceptor ionization energy E_A, the new charge neutrality

[3] The mobile electrons and holes are in equilibrium with the blackbody background radiation modes of photons, and the background thermal blackbody spectra of lattice vibrations, or phonons. In general, the photons in the optical cavity have a temperature T_{photon}, phonons in the crystal cavity have a temperature T_{phonon}, and electrons have a temperature T_{electron}. When the three populations are in thermal equilibrium with each other, they share a common temperature $T = T_{\text{electron}} = T_{\text{phonon}} = T_{\text{photon}}$ that the intrinsic electron and hole distributions depend on. Any non-equilibrium between the three populations is characterized by different temperatures for the three populations: electrons follow the Fermi–Dirac distribution function, whereas photons and phonons follow the Bose–Einstein distribution function.

[4] It is not possible to have a mobile carrier density *lower* than n_i in intrinsic or doped semiconductors with shallow dopant activation energies, setting a minimum conductivity.

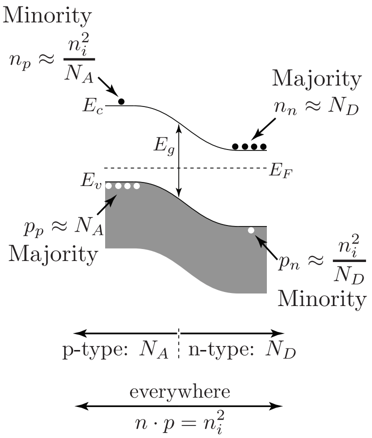


Fig. 16.2 Majority and minority mobile carrier concentrations in the conduction and valence bands at thermal equilibrium. As an example, for silicon at 300 K, $n_i \approx 10^{10}/\text{cm}^3$. Doping of $N_A = N_D = 10^{17}/\text{cm}^3$ results in majority carrier concentrations $n_n = p_p = 10^{17}/\text{cm}^3$ and minority carrier concentrations $n_p = p_n = n_i^2/N_D = 10^3/\text{cm}^3$ on the two neutral sides of the p-n junction outside the depletion regions. Note that the majority carrier concentrations are $\sim 10^{14}$ times larger than the minority carrier concentrations.

condition for a d-dimensional semiconductor is $n_d + N_A^- = p_d + N_D^+$, which is written explicitly as

$$N_c^d F_{\frac{d-2}{2}}\left(\frac{E_F - E_c}{k_b T}\right) + \frac{N_A}{1 + 2e^{\frac{(E_A - E_F)}{k_b T}}} = N_v^d F_{\frac{d-2}{2}}\left(\frac{E_v - E_F}{k_b T}\right) + \frac{N_D}{1 + 2e^{\frac{-(E_D - E_F)}{k_b T}}}. \tag{16.5}$$

The solution to this equation gives the Fermi level E_F and the carrier density explicitly. The number of valence bands considered determines the prefactor of the exponent in the N_A term. It is 2 (as here) for a single valence band, and 4 when both light hole and heavy hole bands are considered. The non Fermi–Dirac statistics of the ionized dopants occurs because of on-site Coulomb repulsion upon occupation of more than one electron at a dopant site, a many-particle effect.

The conduction band edge $E_c(x)$ and the valence band edge $E_v(x)$ are functions of space via the doping, and the choice of semiconductor material via the energy bandgap $E_g(x) = E_c(x) - E_v(x)$. **At thermal equilibrium, the Fermi level is the same everywhere in space.** In non-equilibrium conditions, the quasi-Fermi levels of electrons $F_n(x)$ and holes $F_p(x)$ can vary in space and there may be an energy splitting between them depending on the cause of the non-equilibrium. In the non-degenerate case, the variation of the electron and hole densities in real space can always be written as

$$n(x) = N_c e^{-\left(\frac{E_c(x) - F_c(x)}{k_b T}\right)}, \; p(x) = N_v e^{-\left(\frac{F_v(x) - E_v(x)}{k_b T}\right)}, \text{ and}$$
$$\boxed{n(x)p(x) = n_i^2(x) \cdot e^{\left(\frac{F_c(x) - F_v(x)}{k_b T}\right)}.} \tag{16.6}$$

At thermal equilibrium, $F_c = F_v = E_F$, and $np = n_i^2$ implies if a semiconductor is doped with N_D shallow donors to make it n-type with $n_n \approx N_D$, the minority hole carrier density in the n-type semiconductor is much smaller, but not zero because $n_n \cdot p_n = n_i^2$. Thus, in semiconductors we have the **minority equilibrium carrier concentrations**

$$\boxed{p_n = \frac{n_i^2}{n_n} \approx \frac{n_i^2}{N_D}} \text{ and } \boxed{n_p = \frac{n_i^2}{p_p} \approx \frac{n_i^2}{N_A},} \tag{16.7}$$

and the minority non-equilibrium concentrations are given by Equation 16.6. Fig. 16.2 shows the majority and minority carrier densities in a p-n junction at equilibrium as a representative example. In **Chapter 18** we will use the physics of minority carriers to explain the operation of p-n junctions, bipolar transistors, and electronic gain. Then in **Module III** we will use it heavily to explain photonic devices such as solar cells, LEDs, and lasers.

16.3 Ballistic transport in semiconductors

In **Chapter 5** we had discussed the ballistic transport of *free* electrons – for example see the ballistic current and Fermi levels shown in Fig.

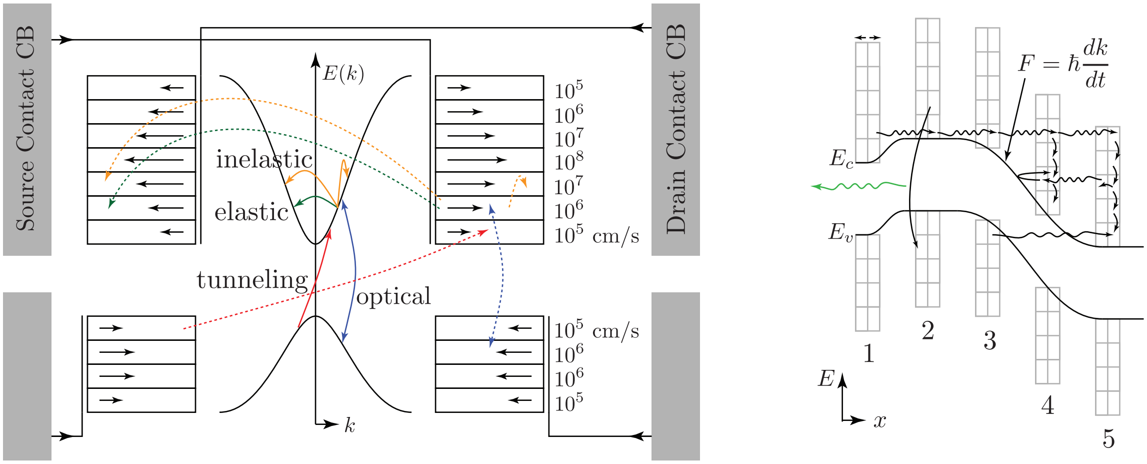


Fig. 16.3 Each k-state in the bandstructure $E(k)$ represents a mode, or a "lane" for electron wave propagation with group velocity $v_g = \frac{1}{\hbar}\frac{dE}{dk}$. Representative velocities are shown with arrows, for both the conduction and valence band. For the same crystal momentum $\hbar k$, the group velocities of states in the conduction band and valence band point in opposite directions. An optical transition occurs between states at the same k that are moving in opposite directions. An interband tunneling transition occurs between states that are moving in the same direction but have different k. These interband processes, as well as the intraband elastic and inelastic scattering events are shown in the bandstructure, and also in the band diagram on the right which highlights how these processes occur in real space. Electrons can propagate ballistically without scattering in a region where the bands bend and there is an electric field, the force $F = \hbar \frac{dk}{dt}$ raises the k and hence the kinetic energy, while maintaining the same total energy. Inelastic scattering events such as phonon emission reduce the energy. Elastic and inelastic scattering can back-scatter electrons, which can then propagate in the opposite direction. Electron waves can reflect (or backscatter) from the band edge. The relative rates of these processes depend on the materials, device geometry, and the driving forces such as a voltage or light incidence.

5.8. We borrowed the free–electron concepts and updated them for semiconductor bands in **Chapter** 8. Let the conduction band of a d-dimensional semiconductor have a fixed electron density[5] of n_d. If the electron states in this band are connected to the electron states in two transparent (= ohmic) metal contacts, and a voltage V is applied across the contacts, the resulting ballistic current is obtained in the following manner. The left electrode is called the source, with a Fermi level E_{Fs}, and the right electrode the drain with Fermi level E_{Fd}, where $E_{Fs} - E_{Fd} = qV$. The electron population of density n_d in the semiconductor conduction band at the "source injection point"[6] is split into right-going states and left going states, indicated in Fig. 16.3. In ballistic transport, there is no scattering – meaning states moving to the right are in equilibrium with the source, and states moving to the left are in equilibrium with the drain. Because the net carrier density is not changed by the applied drain voltage, we have the relation

$$\boxed{n_d = \frac{1}{2} N_c^d \left[F_{\frac{d-2}{2}}\left(\frac{E_{Fs} - E_c}{k_b T}\right) + F_{\frac{d-2}{2}}\left(\frac{(E_{Fs} - qV) - E_c}{k_b T}\right)\right].} \qquad (16.8)$$

Solving this equation at the applied voltage V allows the unknown E_{Fs}, and consequently $E_{Fd} = E_{Fs} - qV$ to be uniquely determined. When $V = 0$, the left-going electron and right-going electron densities

[5] The density may be chemically fixed by doping it with donors, or capacitively fixed by applying a voltage across a gate capacitor.

[6] The flat part of the energy band diagram on the right labeled "2" in Fig. 16.3 may be crudely called the source injection point. In ballistic transport, parts of the electron distribution in the k-space share their Fermi levels with appropriate electrode potentials. In other words, for ballistic transport the electron distribution in the bands is the equilibrium Fermi–Dirac distribution, but with a quasi-Fermi levels that are equal to the Fermi levels of the appropriate electrodes.

are identical, $E_{Fs} = E_{Fd} = E_F$, and the current is zero. For small voltages, E_{Fs} and E_{Fd} split nearly symmetrically around the zero-bias E_F. For large voltages, the argument $\frac{(E_{Fs}-qV)-E_c}{k_bT}$ of the second term which represents the left-going carriers at the source injection point becomes large and negative, meaning $F_{\frac{d-2}{2}}(\eta) \approx e^{\eta}$ becomes very small. In this case, almost all electrons in the band are moving to the right, and almost none to the left. Under this situation, E_{Fs} is in the band, but E_{Fd} enters the gap. For voltages larger than this voltage, E_{Fs} does not change, and the increase in voltage goes into lowering E_{Fd}. This was discussed in detail in Chapter 5.

Because of the imbalance of left- and right-going states when $V \neq 0$, a net current flows in the semiconductor band. To find the net quantum mechanical current, we sum the current due to individual k-states $J_c^d = q\frac{g_s g_v}{L^d}\sum_k f(k)\mathbf{v_g}(\mathbf{k})\cdot\frac{\mathbf{F}}{F}$ where the component of the group velocity along the electric field at the source injection point is used. The sum may be written in the form

$$J_c^d = \frac{q}{L^d}\cdot g_s g_v \cdot [\sum_k f_R(k)\mathbf{v_g}\cdot\hat{z} - \sum_k f_L(k)\mathbf{v_g}\cdot\hat{z}], \tag{16.9}$$

where the dot product gives the projection of the group velocity along the direction of electric field, which is the direction of net current flow. $f_R(k)$ is the Fermi–Dirac distribution of the right-going electron states, characterized by the source Fermi level E_{Fs}, and $f_L(k)$ is the Fermi–Dirac distribution of the left-going electron states, characterized by the drain Fermi level $E_{Fd} = E_{Fs} - qV$. Using integrals of the form

$$\int_0^\infty dk \frac{k^d}{1+e^{\frac{E_c+\frac{\hbar^2k^2}{2m_c^\star}-E_F}{k_bT}}} = I(d,\eta) = \frac{1}{2}\cdot(\frac{2m_c^\star k_bT}{\hbar^2})^{\frac{d+1}{2}}\cdot\Gamma(\frac{d+1}{2})\cdot F_{\frac{d-1}{2}}(\eta), \tag{16.10}$$

where $\eta = \frac{E_F-E_c}{k_bT}$, we get

$$\boxed{J_c^d = \frac{q^2}{h}\cdot N_c^{d-1}\cdot(\frac{k_bT}{q})\cdot[F_{\frac{d-1}{2}}(\frac{E_{Fs}-E_c}{k_bT}) - F_{\frac{d-1}{2}}(\frac{(E_{Fs}-qV)-E_c}{k_bT})].} \tag{16.11}$$

See the table in Fig. A.1 in the Appendix for a summary for the ballistic current in various dimensions. The total electron density of Equation 16.8 is the *sum* of the left- and right-going electron densities, but the total current in Equation 16.11 is the *difference* of the right and left going currents, as it should be. The current density in d-dimensions is proportional to the band-edge DOS of $d-1$ dimensions N_c^{d-1}, and the quantum of conductance $\frac{q^2}{h}$. When the voltage V is large, the left-going current becomes negligible, J_c^d becomes independent of the applied voltage V, and the net current *saturates*. The saturation of the current in a semiconductor – even in the quantum limit – is one of its very important characteristics that enables its useful applications in electronics.

If the electron density in the band is controlled neither by the source or the drain, but **capacitively by a third terminal** called the gate, then n_d of Equation 16.8 can be changed dynamically[7]; E_{Fs} as a result will be controlled by the gate voltage. This allows the current of Equation 16.11 flowing between the source and the drain to be controlled remotely by a third terminal, which does not draw any current itself. In other words, the resistance between the source and drain is "transferred" to a third terminal, giving this device the name the **Tran**sferred-Re**sistor**, or the **Transistor**. Because the control is exerted due to an electric field across a gate capacitor, this form of a transistor is called the **F**ield-**E**ffect **T**ransistor, or the **FET**. Fig. 16.4 shows the physical embodiment, and the electrical characteristics of a typical FET. The expressions in Equation 16.11, suitably adapted for the gate voltage control of E_{Fs} explain all the characteristics of the ballistic FET quantitatively, and is discussed in **Chapter 19**. Scattering in the semiconductor will lower the net current that flows by mixing left and right going states and lowering the effective $E_{Fs} - E_{Fd}$ across the channel from its ballistic value of qV. But it will not change the fundamental qualitative picture described above in the quantum limit. We will apply these concepts to understand quantized conductance and capacitance in **Chapter 17**, Schottky diodes in **Chapter 18**, and ballistic and scattering-dominated FETs in **Chapter 19**.

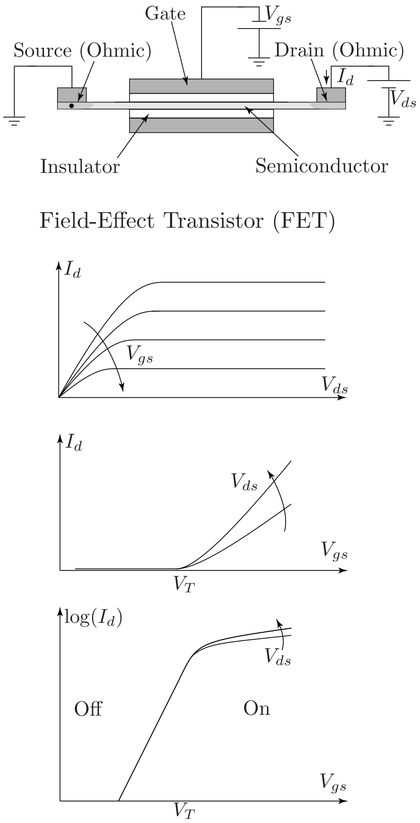


Fig. 16.4 A 2D semiconductor crystal FET channel sandwiched between insulators and metals. The output characteristics of a FET has a saturating drain current I_d vs V_{ds} in the on-state when $V_{gs} > V_T$, and a sharp threshold voltage V_T separating the off and on states.

16.4 Ballistic transport in non-uniform potentials/tunneling

Consider the situation when an electron in the conduction (or valence) band experiences a potential that varies in space according to $E_c(x)$ as shown in Fig. 16.5. The non-uniformity may arise due to band bending due to electric fields, at junctions, or in quantized heterostructures such as quantum wells or dots. Because a Bloch wavefunction is $\psi(x) = e^{ikx}u_k(x)$, the probability density $|\psi(x)|^2 = |u_k(x)|^2$ is periodic, and extends over the entire crystal. Thus, Bloch waves are not well suited to model the quantum state of electrons in non-uniform potentials. However, nearly every electronic and photonic device uses non-uniform potentials to guide electrons and holes. By considering the linear combination of a small set of k states of Bloch waves near the edge of the band, the effective mass approximation converts the Bloch picture of individual electron states $\psi_k(x) = e^{ikx}u_k(x)$ to a wavepacket $\phi(x) = \sum_k C_k\psi_k(x) \approx C(x)u_k(x)$, where $C(x)$ is a slowly varying function compared to the lattice constant. While the Bloch function responds to atoms, the envelope function $C(x)$ responds to variations in the band edges in real space. The resulting effective mass equation for a bandstructure $E_c(x,k) = E_c(x) + \frac{\hbar^2k^2}{2m_c^\star}$, and any perturbation $W(x)$ in the band edge is then

---

[7] The quantitative dependence of a varying n_d on the ballistic current was discussed in Chapter 5. See for example Figures 5.8 for 1D, 5.14 for 2D, and 5.18 for 3D.

$$[E_c(-i\frac{d}{dx}) + W(x)]C(x) \approx EC(x)$$
$$\implies [-\frac{\hbar^2}{2m_c^\star}\frac{d^2}{dx^2} + W(x)]C(x) \approx [E - E_c(x)]C(x), \qquad (16.12)$$

from where the envelope function $C(x)$ is determined. $C(x)$ should not be mistaken for the total wavefunction, which is $\phi(x) \approx C(x)u_k(x)$, but for much phenomena that occur for small energies within the same band, $u_k(x)$ will typically cancel in Equation 16.12 and not play a role. For a constant $W(x) = 0$ and a spatially constant $E_c(x) = E_c$, the envelope function $C(x) = \frac{1}{\sqrt{L}}e^{ikx}$ is a plane wave, where $k = \sqrt{\frac{2m_c^\star(E-E_c)}{\hbar^2}}$. But when the band edge $E_c(x)$ varies slowly for $W(x) = 0$, in **Chapter** 24 we will find that the solution of $C(x)$ for the equation on the right of 16.12 is given by the Wentzel–Kramers–Brillouin or WKB approximation:

$$\frac{d^2C(x)}{dx^2} \approx \underbrace{-\frac{2m_c^\star}{\hbar^2}[E - E_c(x)]}_{Q(x)} C(x) \implies C(x) \approx \frac{K}{Q(x)^{\frac{1}{4}}} e^{\pm\int_a^x du\sqrt{Q(u)}}, \qquad (16.13)$$

where K is a normalization constant. If $E > E_c(x)$ as indicated by the state of energy E_1 in Fig. 16.5, $Q(x) = -k(x)^2 < 0$, and the exponential $e^{\pm\int du\sqrt{Q(u)}}$ in the WKB wavefunction is of the form $e^{\pm i\int dx\cdot k(x)}$, which is still oscillatory, like the plane wave situation. The electron density corresponding to the envelope function (*not* the total wavefunction[8]) is then given by the probability density $n(x) \approx |C(x)|^2 = |K|^2/|k(x)|$, from where it may be seen that for regions where $k(x)$ is large, $n(x)$ is small. The analogy to classical mechanics is the following: a particle spends less time in a region where it is moving fast, i.e., where its kinetic energy is large. The approximate current density $J \approx qnv \approx q(|K|^2/k(x))\cdot(\hbar k(x)/m_c^\star) \approx q\hbar|K|^2/m_c^\star$ ensures the current carried by the wavepacket is continuous in space.

When an electron of energy E lower than the barrier height is incident on a barrier as indicated by the state of energy E_2 in Fig. 16.5, $E < E_c(x)$ on the right side of the region, and $Q(x) = \kappa(x)^2 > 0$. In this case, the WKB envelope function $C(x) \approx \frac{K}{\sqrt{\kappa(x)}}e^{\pm\int dx\cdot\kappa(x)}$, which now is a exponentially decaying, or growing wavefunction amplitude, quite unlike the oscillatory function seen for $E > E_c(x)$. For a thick potential barrier, the electron wave is reflected back after a small, yet finite incursion into the classically forbidden region where $E_2 < E_c$.

If on the other hand the barrier was of a finite thickness between $x_1 < x < x_2$, for propagation of the wavepacket from x_1 to x_2 the WKB transmission probability **through the barrier** is

$$T_{WKB} \approx |\frac{C(x_2)}{C(x_1)}|^2 \approx e^{-2\int_{x_1}^{x_2} dx\cdot\kappa(x)} = e^{-2\int_{x_1}^{x_2} dx\cdot\sqrt{\frac{2m_c^\star}{\hbar^2}[E_c(x)-E]}}. \qquad (16.14)$$

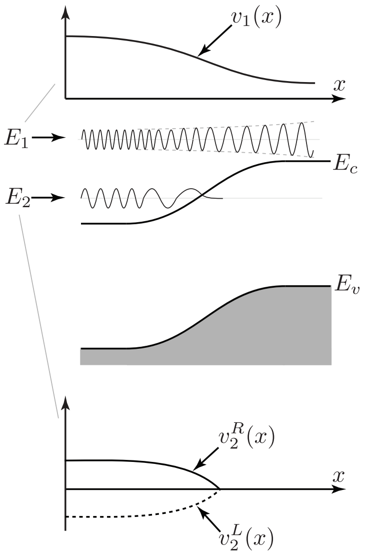


Fig. 16.5 Ballistic transport in non-uniform potentials. If the potential is smoothly varying and the electron energy E_1 is larger than the barrier height, there are no reflections, and the electron wave velocity $v(x)$ decreases and amplitude increases as it moves to the right to maintain current continuity. If the electron energy E_2 is smaller than the barrier height, an electron wave incident from the left with velocity $v_2^R(x)$ penetrates slightly into the barrier, and is completely reflected back with velocity $v_2^L(x)$. If the thickness of the barrier is comparable to the penetration depth of the wavepacket, there is a finite probability of tunneling.

[8] The periodic part of the Bloch function always remains as a constant background oscillation, which does not change the arguments. $u_k(x)$ is similar to the individual compartments of a long train, and $C(x)$ is the train when looked from far – say from an airplane flying at a great height above the train, when the individual compartments are not seen. By looking at $C(x)$, we are looking at the transport of the entire train, or the entire electron wavepacket.

As an example, consider a spatially uniform potential barrier $V_0 - E$, and the total barrier thickness to be $x_2 - x_1 = t_b$. Using them in Equation 24.10, we get $T_{WKB} \approx \exp\,[-\frac{t_b}{0.1\text{ nm}}\sqrt{\frac{m_c^\star}{m_e}\cdot\frac{V_0-E}{1\text{ eV}}}]$. For $V_0 - E = 1$ eV, $m_c^\star = m_e$ and $t_b = 3$ nm, the tunneling probability is $1/e^{30} \approx 10^{-13}$. If the barrier thickness decreases by three times to $t_b = 1$ nm, the tunneling probability increases substantially to $1/e^{10} \approx 5 \times 10^{-5}$. This trick is used in several semiconductor device applications such as in ohmic contacts, and in resonant tunneling diodes to boost the tunneling currents. **Chapter 24** goes deeper into the topics on transmission and tunneling for which a flavor was provided here.

16.5 Scattering of electrons by phonons, defects and photons

The ballistic picture of transport assumes that the distribution functions $f(k)$ of the electrons in the ideal Bloch bands $E_n(k)$ are at the mercy of the contacts to the electron states in the bands. Because of this, the occupation function $f(k)$ of the band electrons shares the same Fermi-level as the contact(s) with which the band can freely exchange electrons. The hidden assumption in this picture is that the presence of non-idealities in the crystal, or perturbations arising in the surroundings of the crystal do not mix the electrons between different k-states against the wishes of the contacts. This is a good approximation in the smallest devices. But it is possible for the physical processes that are not considered in the Bloch bandstructure to wrest part of the control of the occupation function of the electron states from the contacts. These processes are schematically shown in Fig. 16.6. Each of them can be useful, or may create losses. Understanding their nature is critical to coax the desired function from semiconductor devices. In fact each of the processes can be put to use in some form of useful semiconductor device as long as the mechanism is understood.

The processes divide crudely into two types. The first type of processes cause large changes in the energy of the electrons, thereby enabling transitions from one band to another across the bandgap, called **interband** scattering processes. Radiative or photonic transitions, and non-radiative transitions belong to the interband category. The second type of scattering causes minor changes in the energy (inelastic scattering), or no change in the energy (elastic scattering) of the electrons. They leave them within the same band and are called **intraband** scattering processes. Intraband scattering processes change the momentum of electrons, thereby mixing various k-states, and lead to electrical resistivity. Phonon and defect scattering belong to the intraband scattering process[9]. Using Fig. 16.6, we discuss these processes as an appetizer for the more detailed treatment in the following chapters.

Photons [Module IV]: Consider a semiconductor crystal on which photons or light is incident. In **Chapters 26 and 27** of Module IV

[9] A defect that is deep in the bandgap can trap, or bind an electron removing it from the Bloch band energy spectrum and localizing it. We saw an example of this in Chapter 13, Section 13.4. If the deep level interacts with only one band, it captures and releases free carriers, causing fluctuations, or noise in the band current. If a deep level can interact with both bands, it can cause interband recombination. In several following chapters, we will find that processes mediated by deep levels play critical roles in the operation of semiconductor devices.

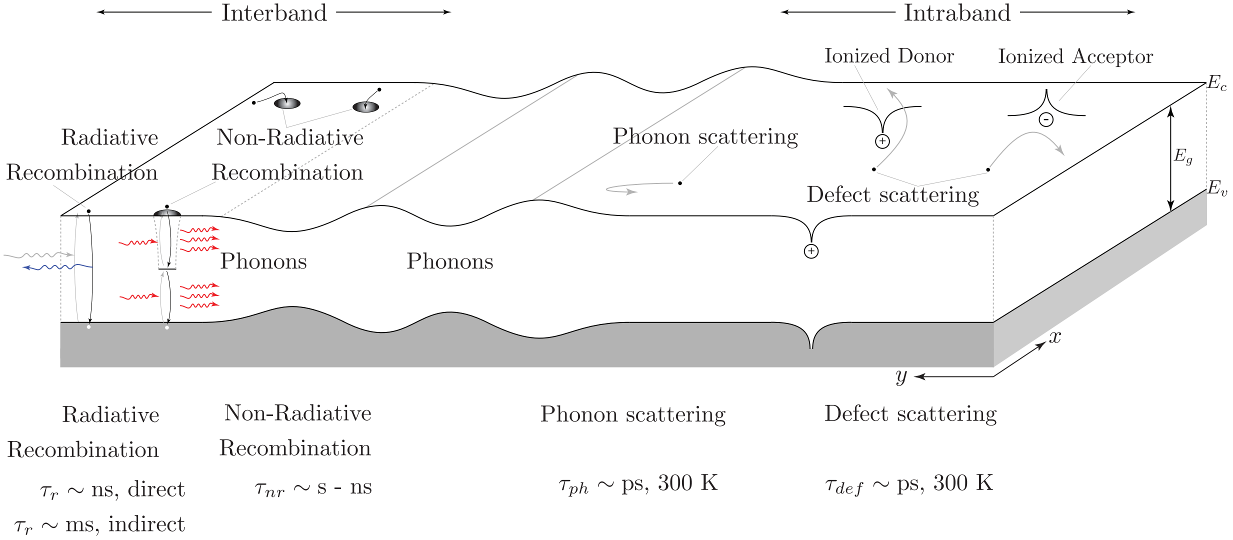

Interband
Intraband
Ionized Donor
Ionized Acceptor
E_c
E_g
E_v
Radiative Recombination
Non-Radiative Recombination
Phonon scattering
Defect scattering
Phonons
Phonons
x
y
Radiative Recombination
Non-Radiative Recombination
Phonon scattering
Defect scattering
$\tau_r \sim$ ns, direct
$\tau_r \sim$ ms, indirect
$\tau_{nr} \sim$ s - ns
$\tau_{ph} \sim$ ps, 300 K
$\tau_{def} \sim$ ps, 300 K

Fig. 16.6 Schematic representation of various scattering processes that mix electrons states within bands, and between bands of a semiconductor. On the left are shown **interband** processes that move electrons from one band to another: the absorption of a photon moves an electron from the valence band to the conduction band, and the emission of a photon causes the reverse process. Absorption therefore creates, or generates an electron in the conduction band *and* a hole in the valence band, and is therefore labeled as a **Generation** or G-process creating an **electron-hole pair**. Similarly, emission of a photon can occur when an electron drops from the conduction band to the valence band, leading to the annihilation of an electron-hole pair, and is referred to as a **Recombination** or R-process. The recombination/generation (R-G) processes may be radiative, or mediated by deep-level states accompanied by emission of phonons, in which case they are called non-radiative, or commonly referred to as Shockley–Read–Hall (SRH) processes. As opposed to the interband R-G processes, scattering mechanisms called **intraband** processes are shown on the right. These processes cause small or no changes to the energy of electrons and leave them in the same band. These mechanisms include scattering of electrons by phonons, and scattering by defects and impurities. Note the typical time scales of the various scattering mechanisms: ranging from ∼ps for intraband processes to ∼ns for radiative processes, and ∼ms-seconds for non-radiative processes.

we review that photons in a light wave are the quanta of electromagnetic excitations of the vacuum. The electric field profile of a light wave is an oscillation in space and time (spacetime) of the form $E(x,t) = 2E_0 \cos(kx - \omega t) = E_0[e^{i(kx-\omega t)} + e^{-i(kx-\omega t)}]/2$, where E_0 is the amplitude of the electric field. The circular frequency ω and the photon wavevector $k = 2\pi/\lambda$ are related by $\omega = ck$ where c is the speed of light and λ the wavelength. The electric potential seen by the electrons in the crystal is the integral of the electric field profile $W(x,t) = \int dx \cdot E(x,t)$. If the light wavelength is of the order of μm, which is much larger than the lattice constant and electron wavelengths in the crystal, the light wave may be approximated as $E(x,t) \approx E_0 e^{\pm i\omega t}$ (this is called the dipole approximation), and the perturbation seen by the electrons in the semiconductor is $W(x,t) = q\int dx \cdot E(x,t) \approx qE_0 x e^{\pm i\omega t} = W_0(x)e^{\pm i\omega t}$, where the perturbation is split into a spatial part $W_0(x)$ and a time-dependent part $e^{\pm i\omega t}$. This potential is experienced by the electrons, in addition to the perfect periodic potential $V_{per}(x)$ of the crystal, so the total potential seen by the electrons is $V_{per}(x) + W(x,t)$. The periodic potential $V_{per}(x)$ creates the band states, and the perturbation due to light $W(x,t)$ causes transitions between them, both upwards for absorption, and downwards for emission as indicated in Fig. 16.6. **Module IV** of the book is dedicated to the study of these photonic interband processes, which typically occur at ∼ns timescales and the devices that result from them such as **light emitting diodes**, **solar cells**, and **lasers**.

Phonons [Chapter 22]: Even when light is not incident on a semiconductor, the atoms in the crystal do not sit still, they vibrate around their equilibrium periodic positions[10]. In **Chapter 22** we will find that the amplitude of their vibrations in space increases with the lattice temperature T. At a high temperature when the vibration amplitude approaches the interatomic distance, the crystal melts. At room temperature, the vibrations from the mean positions are not too large, and can be broken down into quanta of orthogonal lattice wave modes called phonons. The displacement of an atom at location x at time t due to a single phonon is written as $u(x,t) = u_0 e^{i(qx-\omega t)}$. This phonon has an amplitude u_0, a wavevector $q = 2\pi/\lambda$, and frequency ω. In crystalline solids, ω and q have two relations: acoustic, and optical. For acoustic modes, $\omega = v_s q$ where v_s is the sound velocity, and for optical modes, $\omega \approx \omega_0$. The perturbation to the periodic crystal potential $V_{per}(x)$ due to the phonon is treated as a small time-dependent change due to the dilation or compression of the lattice. This is indicated as a greatly exaggerated change to the energy band diagram in Fig. 16.6, where regions of compressive strain have a larger bandgap and dilated regions have a smaller bandgap. A deformation potential D_c is used to measure the change in the electronic potential experienced by the electron: $W(x,t) = D_c \frac{du(x,t)}{dx}$ for a longitudinal acoustic wave. For such a wave, $W(x,t) = iqu_0 e^{i(qx-\omega t)} = W_0(x)e^{\pm i\omega t}$. Note that for both phonons and photons, the perturbation to the periodic potential is of the form $W_0(x)e^{\pm i\omega t}$, and is explicitly time-dependent. The +/- signs in the exponent of the time dependent part of the perturbation $e^{\pm i\omega t}$ lead to the emission and absorption process for both photons and phonons.

Defects [Chapter 23]: Let us say that the crystal is sitting in the dark, and at a very low temperature. Even then, if there are defects in the crystal, they will introduce perturbations to the periodic potential, which remain active also at finite temperatures. Let us say each defect has a potential $W_0(x)$, and there are identical defects at various locations $x_1, x_2, x_3, ...$ Then, the total potential seen by the electron is $V_{per}(x) + W_{tot}(x)$, where $W_{tot}(x) = W_0(x - x_1) + W_0(x - x_2) + W_0(x - x_3) + ...$ Note that the potential of these defects is time-independent[11], unlike those of phonons and photons. Such defect potentials may be considered to be time-dependent, but with $\omega = 0$.

Fermi's golden rule for transitions [Chapter 20]: The electrons in the semiconductor crystal are subject to the above variety of perturbation potentials of the form $W(x,t)$. The time-dependent Schrödinger equation, and the resulting scattering rate of states due to the perturbation derived in **Chapter 20** is given by

---

[10] Imagine we had a very fast camera, and took a snapshot of the crystal. We will then discover that the atoms are actually *never* exactly periodic! So does the entire buildup of the Bloch theory of the electron bandstructure of semiconductors collapse? Luckily, it does not: the phonon energies are typically much smaller than the energy bandgaps, and for most cases it is safe to assume the phonons as *small perturbations* to the perfectly periodic crystal potential. This is the reason why even though at any given time instant the atomic locations are not exactly periodic, we can still observe a strong X-ray diffraction peak as discussed in Chapter 10, and in Exercise 10.6 where the concept of the structure factor, and the Debye–Waller factor was introduced.

[11] The defects that fall in this category of scatterers include ionized donors and acceptors, point defects such as charged or uncharged vacancies and interstitials, linear defects such as dislocations, and planar defects such as grain boundaries. The scattering potentials due to such defects do not change the energy of the electron states in a scattering event, but changes the momentum.

$$i\hbar\frac{\partial\Psi}{\partial t} = [-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2} + V_{per}(x) + W(x,t)]\Psi$$
$$\implies \boxed{\frac{1}{\tau(k \to k')} \approx \frac{2\pi}{\hbar} \cdot |\langle k'|W_0(x)|k\rangle|^2 \cdot \delta[E(k) - E(k') \pm \hbar\omega].} \tag{16.15}$$

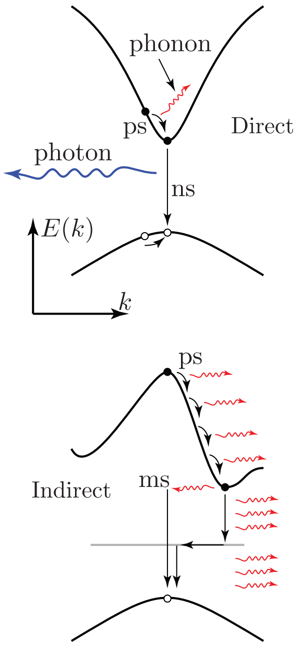


Fig. 16.7 Electronic transitions immediately following the absorption of a photon in a direct gap semiconductor, and an indirect bandgap semiconductor. The direct gap semiconductor results in a radiative transition, whereas the indirect bandgap semiconductor results in a non-radiative transition. The competition between radiative times τ_r and non-radiative times τ_{nr} in a semiconductor defines it's internal quantum efficiency IQE $= \frac{1/\tau_r}{(1/\tau_r)+(1/\tau_{nr})}$. If $\tau_r << \tau_{nr}$, IQE→1 which is the case for direct bandgap semiconductors. This will be discussed in Module IV.

The scattering rate from the unperturbed Bloch state $\psi_k(x) = e^{ikx}u_k(x)$ to the unperturbed Bloch state $\psi_{k'}(x) = e^{ik'x}u_{k'}(x)$ by the perturbation $W(x,t) = W_0(x)e^{\pm i\omega t}$ is given by the Fermi's golden rule. The scattering rate $1/\tau(k \to k')$ is in units of $1/\text{sec}$. The matrix element $\langle k'|W_0(x)|k\rangle = \int dx\psi^\star_{k'}(x)W_0(x)\psi_k(x)$ consists only the spatial part of the perturbation. The time-dependent part of the perturbation appears as the frequency ω inside the Dirac delta function. In the matrix element, the entire Bloch functions $\psi_k(x) = e^{ikx}u_k(x)$ and $\psi_{k'}(x) = e^{ik'x}u_{k'}(x)$ appear. The spatial part $\langle k'|W_0(x)|k\rangle$ dictates whether a scattering transition from state $k \to k'$ is allowed or not, and if it is allowed, it determines its strength. It also ensures momentum conservation in the scattering process. The Dirac delta function ensures energy conservation in the scattering process. For time-independent scattering potentials such as from defects, $\omega = 0$, and the energy of the electron before scattering $E(k)$ and after scattering $E(k')$ must be identical. This form of scattering is called **elastic scattering**. On the other hand, for time-dependent scattering potentials, the electron can gain or lose energy, in processes called **inelastic scattering**. In fact, Fermi's golden rule says more: because $E(k') = E(k) \mp \hbar\omega$, only those transitions are allowed in which the electron can absorb energy from, or emit energy into, phonon and photon modes in discrete *quanta* of energy $\hbar\omega$.

A typical order of magnitude of scattering time for a conduction to valence band (or interband) transition due to electron-photon interaction in a semiconductor is $\tau^{interband}_{optical} \sim$ ns. For transitions within the same band (or intraband transitions) due to phonons or photons, a typical time is $\tau^{intraband}_{photon} \sim \tau^{intraband}_{phonon} \sim$ ps. Because scattering due to defects is typically elastic, it is intraband; the scattering times due to defects can vary widely depending on the defect densities, from $\tau_{defect} \sim$ ms in ultraclean defect-free semiconductors, to $\sim$ ps. As an example, consider the situation just after the creation of an electron hole pair by the absorption of a photon by a direct gap and an indirect gap semiconductor, as shown in Fig. 16.7. Electrons in the conduction band rapidly lose energy to phonons at ps time scale, and relax to the bottom of the band, and holes in the valence band emit phonons to bubble up to the top. They then wait for about $\tau_r \sim$ns in the direct-gap semiconductor to recombine and emit a photon. On the other hand, for an indirect bandgap semiconductor, the radiative recombination lifetime is far too long; all the energy of the photon is dissipated as heat by the emission of phonons.

16.6 The Boltzmann transport equation

Because there are a large number of electron states in semiconductor bands and the several quantum processes discussed occur simultaneously, an accurate book-keeping tool is necessary to understand experimentally measured phenomena. The Boltzmann transport equation (BTE) is that tool. **Chapter 21** is devoted to the details, here we provide an extended preview. We will also realize that the BTE introduces and enforces statistical mechanics into the quantum mechanical processes discussed earlier.

If we can determine the final distribution function $f(k)$ of the electron states considering transitions and scattering between all $k \leftrightarrow k'$, we can find how a semiconductor will respond to voltages, temperature gradients (heat), or to light. For example, the net charge current flowing in the semiconductor then will be $J = q\frac{g_s g_v}{L^d}\sum_k v_g(k) f(k)$. To include the variations of the distribution function in space x and time t, we write the occupation function as $f(x,k,t)$. Though the spatial location of a *Bloch state* of wavefunction $\psi_k(x) = e^{ikx}u_k(x)$ is ill defined because the state extends over the entire crystal, the spatial extent of a *wavepacket* $\phi(x) \approx C(x)u_k(k)$ is well defined by $\Delta x \sim 1/\Delta k$, where the wavepacket is constructed by a linear combination of Bloch states with wavevectors $k \pm \Delta k$. So when we write the distribution function as $f(x,k,t)$, we implicitly mean the occupation function of the wavepacket states, with x meaning $\langle x \rangle$ the mean x of the wavepacket, and k meaning $\langle k \rangle$, the mean k of the wavepacket.

Consider the distribution function $f(x,k,t)$ which is the occupation probability of an electron state k at spatial location x at time t. Let $S(k \to k')$ denote the scattering rate in 1/sec out of state k into other states k'. $S(k' \to k)$ then denotes the reverse, or rate of scattering in to state k from other states k'. These rates are obtained from the Fermi's golden rule result in Equation 16.15. It may be weighted if the physical process so demands; for example, for momentum scattering rate, $S_m(k \to k') = \frac{1}{\tau(k \to k')}[1 - \cos(\theta)]$, where $\frac{1}{\tau(k \to k')} = \frac{2\pi}{\hbar} \cdot |\langle k'|W_0(x)|k\rangle|^2 \cdot \delta[E(k) - E(k') \pm \hbar\omega]$, and θ is the angle between k and k'.

There is an interesting relation between the direction of a scattering event $S(k \to k')$ and its reverse $S(k' \to k)$. Single-particle quantum mechanics in the form of Fermi's golden rule (**Chapter 20**) does not distinguish between a rate and its reverse. But the quantum statistics of the states, a genuine many-particle effect, clearly dictates which processes are preferred. To see this, consider a situation when there are no electric fields, or no concentration gradients in the semiconductor, and $f(x,k,t)$ is the occupation function of an electron wavepacket. Because electrons are fermions, if a state k is already filled, the Pauli exclusion principle kicks in, and it is not possible to scatter in to it. For an electron to scatter from state k' to k, state k' must be occupied and state k must be empty. These conditions dictate that the rate of scattering in to a state k is proportional to the factor $f_{k'}(1-f_k)$, leading to the following rate equation for fermions:

$$\frac{df(x,k,t)}{dt} = \sum_{k'}[S(k' \to k)f_{k'}(1-f_k) - S(k \to k')f_k(1-f_{k'})]. \tag{16.16}$$

At equilibrium, $f(x,k,t)$ does not change with time, and the left hand of this equation is zero. But equilibrium is much more. If the total number of particles does not change with time, the rate of each microscopic scattering process going one way $k \to k'$ and its reverse, $k' \to k$ are also exactly equal. The occupation function under this equilibrium condition is the Fermi–Dirac distribution function for electrons, $f_k^0 = \frac{1}{1+e^{\frac{E(k)-E_F}{k_bT}}}$. Here, E_F is the Fermi level, and T is the temperature of the heat bath with which the semiconductor electrons are in equilibrium[12]. Regardless of the temperature of the electrons in equilibrium, we must then have for all k and k',

$$S(k' \to k)f_{k'}^0(1-f_k^0) - S(k \to k')f_k^0(1-f_{k'}^0) = 0$$
$$\implies \frac{1-f_{k'}^0}{f_{k'}^0} \cdot \frac{f_k^0}{1-f_k^0} = \frac{S(k' \to k)}{S(k \to k')} = e^{\frac{E(k')-E(k)}{k_bT}}. \tag{16.17}$$

For elastic scattering, $E(k) = E(k')$, and $S(k \to k') = S(k' \to k)$. For emission or absorption of photons or phonons, the rate of absorption is proportional to the number of photons or phonons that are available, given by the Bose–Einstein distribution $N(\omega) = 1/[e^{\frac{\hbar\omega}{k_bT}} - 1]$. Thus, $S_{abs} = S[E(k) \to E(k) + \hbar\omega] \propto N(\omega) = 1/[e^{\frac{\hbar\omega}{k_bT}} - 1]$. Then, we obtain

$$\frac{S_{em}}{S_{abs}} = e^{\frac{\hbar\omega}{k_bT}} \implies S_{em} = e^{\frac{\hbar\omega}{k_bT}} S_{abs} \propto e^{\frac{\hbar\omega}{k_bT}} N(\omega)$$
$$e^{\frac{\hbar\omega}{k_bT}} N(\omega) = 1 + \frac{1}{e^{\frac{\hbar\omega}{k_bT}} - 1} = 1 + N(\omega)$$
$$\implies \boxed{S_{em} \propto [\underbrace{1}_{\text{spontaneous}} + \underbrace{N(\omega)}_{\text{stimulated}}] \text{ and } S_{abs} \propto N(\omega).} \tag{16.18}$$

Absorption of course is naturally *stimulated*, meaning if there are no photons or phonons to absorb, $N(\omega) = 0$, and therefore $S_{ab} = 0$. But we see that if an electron is emitting into bosonic modes of photons or phonons, $S_{em} \propto [1 + N(\omega)]$, which means that even if $N(\omega) = 0$ and there are no photons or phonons present, the electron can still emit: this is called *spontaneous* emission. If there are photons or phonons present, the rate of emission is *enhanced*, this is *stimulated* emission, as shown in Fig. 16.8. The possibility of stimulated emission was first conceived by Einstein, and has led to the realization of lasers. By considering a rate equation of a similar kind as Equation 16.16 but for bosons, with f_k replaced by $N(\omega)$ for the equilibrium Bose–Einstein distribution, it may again be shown that $\frac{S(\hbar\omega' \to \hbar\omega)}{S(\hbar\omega \to \hbar\omega')} = e^{\frac{\hbar\omega' - \hbar\omega}{k_bT}}$, which is the same as the relation for fermions (see Chapter 29).

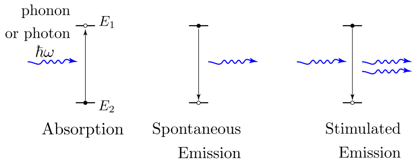


Fig. 16.8 Absorption, Spontaneous Emission, and Stimulated Emission processes. The energy levels E_1 and E_2 could be electron energies in an atom, or a crystal. A photon produced by stimulated emission is identical in every aspect to the one that stimulated it, it is a *clone*.

---

[12] What exactly is this "heat bath"? The atoms of a semiconductor crystal sitting on a chip at room temperature are in thermal equilibrium with the chip, and the electrons in the semiconductor are in thermal equilibrium with the atoms of the crystal. Thus, the temperature of the chip makes it all the way to the temperature characterizing the Fermi–Dirac distribution of the electrons. A theorem of statistical mechanics guarantees that **the equilibrium state of any system is *independent* of the mechanisms that allow the system to attain equilibrium**. For example, if the electrons and holes in a semiconductor were interacting strongly with light in the form of photons from the sun in a solar cell, the equilibrium temperature of the electrons and holes will be somewhere between the temperature of the sun, and the temperature of the solar panel on which the semiconductor sits. What is important is that once any equilibrium T is reached, the statistics are the same regardless of how the carriers were generated, by heat or by light.

Now if we turn on an electric field that exerts a force F on the electrons, it will change the electron wavevector k according to $F = \hbar \frac{dk}{dt}$. Because it will move electrons from one k-state to another, it will create gradients of the occupation function f in the k space. The electron concentration can also change in real space, leading to gradients of f in x. In the presence of these gradients, the distribution function deviates from the equilibrium Fermi–Dirac function, $f_k \neq f_k^0$. If these processes of gradients are considered, Equation 16.16 must change to

$$f(x,k,t) = f(x - v \cdot dt, k - \frac{F}{\hbar} \cdot dt, t - dt) + \\ dt \cdot \sum_{k'} [S(k' \to k) f_{k'}(1 - f_k) - S(k \to k') f_k (1 - f_{k'})], \tag{16.19}$$

which upon Taylor expansion of the first term on the right becomes

$$\boxed{\frac{\partial f_k}{\partial t} + v \frac{\partial f_k}{\partial x} + \frac{F}{\hbar} \frac{\partial f_k}{\partial k} = \sum_{k'} [S(k' \to k) f_{k'}(1 - f_k) - S(k \to k') f_k (1 - f_{k'})].} \tag{16.20}$$

This is the Boltzmann transport equation. The mixing of states $k \leftrightarrow k'$ which was not allowed in our ballistic treatment earlier is now allowed, as seen explicitly in the right hand side of the equation. By solving this equation, we can obtain f_k, the distribution function for electrons, considering all imaginable processes shown schematically in Figures 16.3 and 16.6. Under ballistic conditions, the RHS which captures all forms of scattering may be set to zero. When scattering is present, the right hand side must be included, and the equation solved for f_k. In addition to the steady state behavior, the Boltzmann equation also provides the time-dependent changes of f_k, meaning it captures the dynamical behavior of the electrons in response to time-varying perturbations.

As an illustrative example, consider a situation when there are no concentration gradients $\frac{\partial f_k}{\partial x} = 0$, and no electric fields, $F = 0$. Say the scattering sum on the RHS in Equation 16.20 can be written as a small change from the equilibrium Fermi–Dirac distribution, $RHS \approx -\frac{f_k - f_k^0}{\tau}$ characterized by a scattering time τ. Then, the Boltzmann transport equation simplifies to

$$\frac{\partial f_k}{\partial t} \approx -\frac{f_k - f_k^0}{\tau} \implies f_k(t) \approx f_k^0 + (f_{ss} - f_k^0) e^{-\frac{t}{\tau}}. \tag{16.21}$$

This example shows that if at time $t = 0$, the distribution function was in non-equilibrium, yet steady-state value $f(t = 0) = f_{ss}$, upon removal of all fields and gradients, the distribution function relaxes to the equilibrium Fermi–Dirac value exponentially with a characteristic time τ. Fig. 16.9 shows this relaxation process when the force $F(t)$ drops to zero. If all scattering events are physically captured by $RHS \approx -\frac{f_k - f_k^0}{\tau}$, we call the case the relaxation time approximation, meaning it can be characterized by one time constant τ.

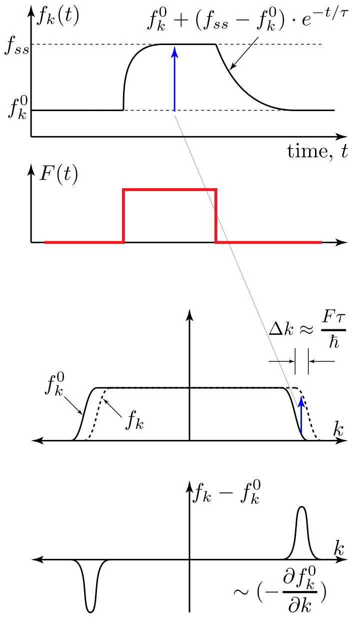


Fig. 16.9 The distribution function as a function of time and resolved in the k-space. The distribution is pulled out from the Fermi–Dirac equilibrium function f_k^0 by an electric field induced force $F(t)$ to a steady state distribution $f_{ss} = f_k$. When the force is removed, the distribution relaxes back to the Fermi–Dirac function with a characteristic relaxation time τ. In the k-space, the distribution function changes from the Fermi–Dirac distribution f_k^0 which is symmetric about $k = 0$ to the steady-state function f_k, which is not symmetric. The shift of f is roughly $\Delta k \approx F\tau/\hbar$ along the k-axis, where F is the force and τ the scattering time. The difference between the steady state and the equilibrium distributions $(f_k - f_k^0)$ is proportional to $-\partial f_k^0/\partial k$ and peaks near the Fermi surfaces. The net electrical current is carried by the distribution that is not symmetric, by states that are within these regions where $(f_k - f_k^0) \neq 0$.

If on the other hand, when the electric field induced force $F(t)$ is turned on from zero as shown in Fig. 16.9, the distribution reaches steady state after some time. When steady state is attained, the distribution function $f_{ss} = f_k$ does not change with time any more, and $\frac{\partial f_k}{\partial t} = 0$. Fig. 16.9 shows the functions in the k-space at equilibrium before the field is turned on (solid line), and at steady state after the field is on for some time (dashed line). Thus the Boltzmann transport equation becomes

$$\frac{F}{\hbar}\frac{\partial f_k}{\partial k} \approx -\frac{f_k - f_k^0}{\tau} \implies f_k \approx f_k^0 - \frac{F}{\hbar}\tau\frac{\partial f_k}{\partial k} = f_k^0 - qEv_k\tau\frac{\partial f_k}{\partial E(k)}$$
$$\implies \boxed{f_k \approx f_k^0 + qEv_k\tau(-\frac{\partial f_k^0}{\partial E(k)})}, \tag{16.22}$$

where we have used the group velocity of the wavepacket $v_k = \frac{1}{\hbar}\frac{\partial E(k)}{\partial k}$, and that the force on the electron is $F = qE$, were E is the electric field. We have further replaced $f_k \to f_k^0$ inside the derivative: this approximation holds for small electric fields. This solution of the Boltzmann transport equation is in a suggestive form because the equilibrium portion f_k^0 is split from the non-equilibrium portion $qEv_k\tau(-\frac{\partial f_k^0}{\partial E(k)})$. The non-equilibrium portion is driven by the electric field E, and is the starting point for the evaluation of electrical conductivity, mobilities, and electric currents in the presence of scattering. Equation 16.22 is the starting point to understand scattering dominated transport called **drift**.

If the electric field was zero, but there was a concentration gradient in the real space, the steady state distribution function in the relaxation time approximation is

$$v_k\frac{\partial f_k}{\partial x} \approx -\frac{f_k - f_k^0}{\tau} \implies \boxed{f_k \approx f_k^0 + v_k\tau(-\frac{\partial f_k^0}{\partial x})}, \tag{16.23}$$

where we have again replaced the unknown $f_k \to f_k^0$ on the right for small gradients compared to equilibrium. Equation 16.23 is the starting point to understand scattering dominated transport called **diffusion**.

16.7 Current flow with scattering: drift and diffusion

Fig. 16.10 schematically represents the processes of drift and diffusion of electrons and holes in semiconductors as space-time diagrams. Electrons are represented as points in space and time. At a finite temperature, the vibrations of the atoms of the semiconductor jostle electrons and holes around. If by some means we inject carriers in a particular position in real space with nearly zero velocity (say at k-states near

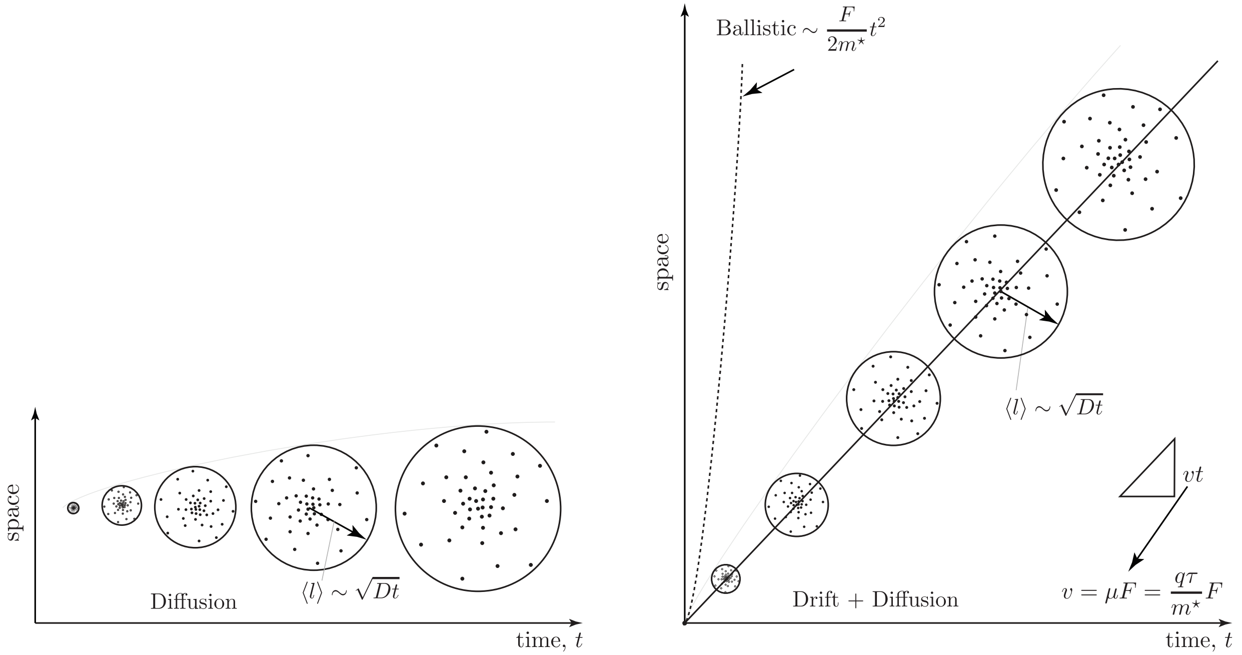

space
Diffusion
$\langle l \rangle \sim \sqrt{Dt}$
time, t
Ballistic $\sim \frac{F}{2m^\star}t^2$
space
$\langle l \rangle \sim \sqrt{Dt}$
vt
Drift + Diffusion
$v = \mu F = \frac{q\tau}{m^\star}F$
time, t

Fig. 16.10 Diffusion of particles in a space-time diagram. The particle (e.g. an electron) is depicted as a dot. The spatial spread $\langle l \rangle$ of a bunch of particles increases as $\sqrt{Dt}$, where t is time, and D is the diffusion constant, which for semiconductors is expressed in units of cm^2/s. Note that the radius does not increase linearly with time, but as $\sqrt{t}$. The figure on the right shows the space-time diagram when a constant force causes the particles to drift, in addition to diffusion. The drift part results in a constant velocity $v = \mu F$ where μ is the mobility in units of cm^2/V·s. At very short time scales, diffusion can cause a net particle flow against the direction of drift. If scattering events were absent, the electrons move ballistically; in the case of a parabolic bandstructure the distance changes as the square of the time.

the bottom of the conduction band), the carriers will not stay there, but will be scattered to various other k-states by the lattice vibrations, or by defects and impurity scattering. As a result of these scattering events, in real space, the carriers perform a random walk characterized by the diffusion equation[13].

In the presence of spatial gradients of the distribution function (or equivalently, gradients in real space in the carrier concentration), the current due to the diffusion of particles is obtained from Equation 16.23. If the momentum scattering time is $\tau_m(k)$ for each k-state, using the usual prescription for quantum mechanical current for any crystal, the diffusion current is:

$$J_{diff} = q\frac{g_s g_v}{L^d}\sum_k v_k[f_k^0 - v_k \tau_m(k)\frac{\partial f_k^0}{\partial x}]$$

$$\implies \boxed{J_{diff} \approx q\frac{g_s g_v}{L^d}\sum_k v_k^2 \tau_m(k)(-\frac{\partial f_k^0}{\partial x}) \approx qD\frac{dn(x)}{dx}.} \qquad (16.24)$$

When simplified, the diffusion current takes the form $J_{diff} = qD\frac{dn(x)}{dx}$,

[13] The diffusion equation, which has a long history, applies to gases, atoms and molecules, and electrons and holes. It also applies to "abstract" quantities such as ideas and rumors! The microscopic models of diffusion track back to the experimental observation of Brownian motion, and Einstein's formulation of the random walk problem.

where $n(x) \sim \frac{g_s g_v}{L^d} \sum_k f_k^0$ is the local electron concentration, and $D \sim \langle v(k)^2 \tau(k) \rangle$ is the diffusion constant that captures the driving force causing the current to flow. Just as diffusion of charge, a spin diffusion current $J_{spin} = \frac{\hbar}{2} D_s \frac{ds(x)}{dx}$ flows in response to gradients in the spin density $s(x)$ with a corresponding spin diffusion constant D_s and spin angular momentum $\hbar/2$. Heat currents $J_{heat} = -\kappa_e \frac{dT(x)}{dx}$ flow in response to temperature gradients. The thermal diffusion constant κ_e here is the electronic contribution to the thermal conductivity, which was introduced in Chapter 2 in the context of the Wiedemann-Franz law.

Fig. 16.10 on the right shows the space-time diagram of particles when a drift current adds to the diffusion. If we neglect diffusion, the presence of an electric field induced force F leads to the Drude-like velocity-field relationship in semiconductor of the form $J_{drift} = qn\mu E$, as discussed in several previous parts of the book. The end result of drift motion in the presence of scattering was already discussed in Chapter 2 for classical electrons, and Chapter 9 for Bloch electrons in crystals satisfying the wave-particle duality of quantum mechanics. At a low electric field E, a terminal or drift velocity $v = \mu E$ is reached, where $\mu = q\tau/m^\star$ is the particle mobility. τ is a characteristic scattering time, and $m^\star$ is the particle effective mass. A constant drift velocity implies a linear increase in the distance traveled in time, $x \sim vt$. When superimposed on the diffusive increase of the radii in time $\langle l \rangle \sim \sqrt{Dt}$, the space-time diagram evolves as indicated. Compare the scattering dominated drift-diffusion space-time trajectory to the ballistic trajectory $x \sim Ft^2/(2m^\star)$ shown in Fig. 16.10, which goes as the square of the time for a parabolic bandstructure.

The Boltzmann equation enables us to connect the microscopic details in the following way. Let the momentum scattering time for a state k be $\tau_m(k)$. In response to an electric potential gradient $E = -\frac{dV(x)}{dx}$ due to an electric field, the charge drift current density is obtained from Equation 16.22 as

$$J_{drift} = q\frac{g_s g_v}{L^d} \sum_k v_k f_k \approx q\frac{g_s g_v}{L^d} \sum_k v_k [f_k^0 + qEv_k \tau_m(k)(-\frac{\partial f_k^0}{\partial E(k)})]$$

$$\implies \boxed{J_{drift} = \frac{q^2 g_s g_v}{L^d}[\sum_k v_k^2 \tau_m(k)(-\frac{\partial f_k^0}{\partial E(k)})]E \approx \underbrace{qn\mu}_{\sigma} E.} \tag{16.25}$$

Here we have used $\sum_k v_k f_k^0 = 0$. For charge current, the momentum relaxation time $\tau_m(k)$ is obtained by first finding $1/\tau(k \to k')$ from the Fermi's golden rule considering individual scattering potentials $W(x,t)$. The momentum scattering rate is obtained by summing the rate of a single state $1/\tau(k \to k')$ over all final states k' with an angular weight for momentum scattering: $\frac{1}{\tau_m(k)} = \sum_{k'} \frac{1}{\tau(k \to k')}[1 - \cos\theta]$, where θ is the angle between k and k', or the angle by which the momentum is scattered. If the angle by which the electron is scattered is small, the net momentum in the direction of current is not affected much, and the conductivity is high. But if the angle is large, say $\theta \approx \pi$, the conductivity is lowered significantly. This is captured by the $1 - \cos\theta$ weight heuristically. This factor can be derived exactly from the Boltzmann transport equation. We will discuss the Fermi's golden rule in **Chapter 20**, and apply it to understand the microscopic details of scattering by phonons in **Chapter 22**, and defects in **Chapter 23**.

For completeness and future reference, we discuss the **continuity equation** and its solution here. The continuity equation is really a statement of particle conservation: similar to its quantum counterpart discussed in Chapter 3. As an example, for electrons in the conduction band of local density $n(r)$ the continuity equation relates the rates of change in the self-explanatory form

$$\underbrace{\frac{\partial n}{\partial t}}_{\text{change in density}} = \underbrace{-\nabla \cdot (\frac{J_n}{-q})}_{\text{density inflow}} + (\underbrace{G_n}_{\text{generation}} - \underbrace{R_n}_{\text{recombination}}), \tag{16.26}$$

where G_n is a generation rate of electrons into the conduction band, and R_n is the recombination (or loss) rate of electrons from the conduction band, and $J_n = qn\mu_n F + qD_n \nabla n$ is the sum of the drift and diffusion current densities where F is the electric field. Consider a 1D example that will be relevant for semiconductor devices in later chapters. A pulse of light creates N excess electrons in the conduction band in a n-type semiconductor of equilibrium density n_0, at $t = 0$ and $x = 0$, in the presence of a constant electric field F. How does the distribution evolve in space and time? To find this, we substitute the expression for current in Equation 16.26. Since at $t > 0$ there is no light, there is no more generation, implying $G_n = 0$. Excess electrons will be lost by recombination at the rate $R_n = (n - n_0)/\tau_n$, where τ_n is the recombination rate. The continuity equation then becomes

$$\frac{\partial(n-n_0)}{\partial t} = D_n \frac{d^2(n-n_0)}{dx^2} + \mu_n F \frac{d(n-n_0)}{dx} - \frac{(n-n_0)}{\tau_n}$$

$$\implies \boxed{n(x,t) - n_0 = \frac{N}{\sqrt{4\pi D_n t}} \cdot \exp\Big(-\frac{(x-\mu_n F t)^2}{4D_n t} - \frac{t}{\tau_n}\Big),} \tag{16.27}$$

where the solution of the partial differential equation takes the form of a Gaussian distribution, whose mean position drifts with the field with drift velocity $v = \mu_n F$ and spread diffuses over length $\sqrt{D_n t}$ in time t, while losing the particle density as e^{-t/τ_n}. The coefficient indicates that the 1D density decreases as $N/\sqrt{4\pi D_n t}$; in d-dimensions, the coefficient is $N/(4\pi D_n t)^{d/2}$. This drift and diffusion phenomena is illustrated in Fig. 16.11.

Because both diffusion and drift processes are linked to the exact same microscopic scattering mechanisms, we can expect the drift current $J_{drift} = qn\mu_n E$ and diffusion current $J_{diff} = qD_n dn/dx$ to be

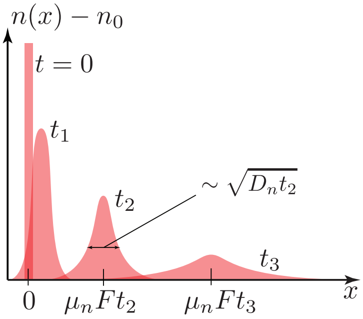


Fig. 16.11 Diffusion, drift, and recombination of excess electrons after they are generated by a pulse of light in a semiconductor. This behavior was verified by the **Haynes-Shockley** experiment, a classic in semiconductor device physics. This measurement enables a measurement of the diffusion constant D_n and the mobility μ_n simultaneously, and helps verify the Einstein relation $D_n/\mu_n = k_b T/q$ derived in Equation 16.28.

linked as well. Consider a semiconductor that has an intrinsic concentration gradient due to say a varying doping in space. In that case, as carriers diffuse, they set up an internal electric field, and at equilibrium the net current must be zero. The electron density depends on the local potential as $n(x) = N_c(x)e^{-(E_c(x)-E_F)/k_bT} = N_c \cdot e^{+qV(x)/k_bT}$, where we write $E_c(x) - E_F = -qV(x)$. Though $N_c(x)$ can vary in space, it is much slower than the exponential term, which is why we can assume it to be nearly constant. Thus we can write

$$J = J_{drift} + J_{diff} = qn\mu_n(-\frac{dV}{dx}) + qD_n\frac{dn}{dx} = 0 \implies \frac{D_n}{\mu_n} = \frac{dV}{dn/n}$$
$$n = N_c e^{\frac{qV}{k_bT}} \implies \frac{dV}{dn/n} = \frac{k_bT}{q} \implies \boxed{\frac{D_n}{\mu_n} = \frac{k_bT}{q}.} \tag{16.28}$$

The boxed part is the **Einstein relation**, which highlights the deep connection between the diffusion and drift processes.

Returning to drift transport, by relating the drift charge current to a Drude form $J_{drift} = \sigma E = qn\mu E$, where $n = \frac{g_s g_v}{L^d}\sum_k f_k$, we identify the electron drift mobility as

$$\mu = \frac{J_{drift}}{qnE} = \frac{\frac{q^2 g_s g_v}{L^d}\cdot[\sum_k v_k^2\tau_m(k)(-\frac{\partial f_k^0}{\partial E(k)})]\cdot E}{q\cdot[\frac{g_s g_v}{L^d}\sum_k f_k]\cdot E} = q\cdot\frac{\sum_k v_k^2\tau_m(k)(-\frac{\partial f_k^0}{\partial E(k)})}{\sum_k f_k^0}. \tag{16.29}$$

The numerator has the derivative of the Fermi–Dirac distribution with energy. This means if we have a degenerate Fermi gas when say $E_F > E_c$, for most of the electrons in the window $E_c \leq E(k) \leq E_F$, the term $-\frac{\partial f_k^0}{\partial E(k)} \approx 0$, and they cannot carry *net* currents. Their contribution to the ensemble mobility is low: this is Pauli blocking in action. Because $-\frac{\partial f_k^0}{\partial E(k)}$ peaks at the Fermi level, the contribution to the mobility will be dominated by states near the Fermi level $E \approx E_F$, or in other words, the conductivity is dominated by states near the Fermi level. If we consider the low temperature approximation $-\frac{\partial f_k^0}{\partial E(k)} \approx \delta(E - E_F)$ in the degenerate condition for parabolic bands, we can obtain $\mu \approx \frac{q\tau_m(k_F)}{m_c^\star}$, where k_F is the Fermi wavevector. This expression of the mobility is in the Drude form. For the more general situation, the summation over the k states must be performed, which we discuss in the next section[14].

16.8 Explicit calculations of scattering rates and mobility

We now succinctly outline the explicit calculation of the drift mobility due to scattering between effective mass wavepacket states centered at states $|\mathbf{k}\rangle$ and $|\mathbf{k}'\rangle$ within the *same band*, with corresponding

[14] We will build upon this brief introduction of transport phenomena in **Chapters 21, 22, and 23**. The basic introduction of the drift and diffusion processed in this chapter will enable us to treat various phenomena that occur in devices such as Schottky and p-n junction diodes in **Chapter 18** without worrying about all the microscopic details.

time-dependent wavefunctions $\phi_{\mathbf{k}}(\mathbf{r},t) \approx C_{\mathbf{k}}(\mathbf{r},t)u_{\mathbf{k}}(\mathbf{r})$ and $\phi_{\mathbf{k}'}(\mathbf{r},t) \approx C_{\mathbf{k}'}(\mathbf{r},t)u_{\mathbf{k}'}(\mathbf{r})$. Consider a scattering potential of the form $W(\mathbf{r},t) = W_0(\mathbf{r})e^{\pm i\omega t}$. The time-dependent effective mass equation for the electron wavepacket centered at $\mathbf{k}$ is

$$\begin{aligned} i\hbar\frac{\partial C_{\mathbf{k}}(\mathbf{r},t)}{\partial t} &= [E_c(-i\nabla) + W(\mathbf{r},t)]C_{\mathbf{k}}(\mathbf{r},t) \\ &= [-\frac{\hbar^2}{2m_c^\star}\nabla^2 + E_c(\mathbf{r}) + W(\mathbf{r},t)]C_{\mathbf{k}}(\mathbf{r},t). \end{aligned} \tag{16.30}$$

The lattice periodic part $u_{\mathbf{k}}(\mathbf{r})$ cancels in this equation, just as in the time-independent version. In the absence of the perturbation, the envelope functions of the states are $C_{\mathbf{k}}(\mathbf{r})$ and $C_{\mathbf{k}'}(\mathbf{r})$. We can write the momentum scattering rate using Fermi's golden rule for the envelope function as

$$\frac{1}{\tau(\mathbf{k}\to\mathbf{k}')} = \frac{2\pi}{\hbar}|\langle C_{\mathbf{k}'}(\mathbf{r})|W_0(r)|C_{\mathbf{k}}(\mathbf{r})\rangle|^2\delta[E(\mathbf{k}') - (E(\mathbf{k}) \pm \hbar\omega)]$$

$$\Longrightarrow \boxed{\frac{1}{\tau_m(\mathbf{k})} = \sum_{\mathbf{k}'}\frac{1}{\tau(\mathbf{k}\to\mathbf{k}')}(1-\cos\theta),} \tag{16.31}$$

where $\cos\theta = \frac{\mathbf{k}'\cdot\mathbf{k}}{|\mathbf{k}||\mathbf{k}'|}$. It is possible to directly use the envelope function in the golden rule for transitions *within the same band*, for small changes in $\mathbf{k}$, which ensures that the periodic part of the Bloch functions $u_{\mathbf{k}}(\mathbf{r})$ are the same. For such transitions, the time-dependent effective mass equation Equation 16.30 has an identical mathematical form as the exact time-dependent equation for the Bloch functions in Equation 16.15. Whenever this condition fails, one must resort back to using the functions $C_{\mathbf{k}}(\mathbf{r})u_k(\mathbf{r})$ and $C_{\mathbf{k}'}(\mathbf{r})u_{k'}(\mathbf{r})$, where the lattice-periodic parts are explicitly included[15].

For transitions within the same band and for small changes in $\mathbf{k}$, consider the case of scattering by a *single* positively charged ionized impurity (say an ionized donor atom) in a semiconductor crystal. Fixing the origin of coordinates at the location of the impurity, the screened scattering potential[16] is $W(r,t) = W_0(r)e^{\pm i\omega t} = W_0(r) = -\frac{q^2}{4\pi\epsilon_s r}e^{-\frac{r}{L_D}}$, where ϵ_s is the dielectric constant of the semiconductor, $L_D = \sqrt{\frac{\epsilon_s k_b T}{q^2 n}}$ is the Debye screening length with n the free–electron concentration, and $\omega = 0$, indicating the scattering potential is time-independent. The envelope functions are $C_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{V}}e^{i\mathbf{k}\cdot\mathbf{r}}$ and $C_{\mathbf{k}'}(\mathbf{r}) = \frac{1}{\sqrt{V}}e^{i\mathbf{k}'\cdot\mathbf{r}}$ where V is the macroscopic volume of the semiconductor crystal. The scattering matrix element is

$$\begin{aligned} \langle C_{\mathbf{k}'}(\mathbf{r})|W_0(r)|C_{\mathbf{k}}(\mathbf{r})\rangle &= \int d^3r(\frac{1}{\sqrt{V}}e^{-i\mathbf{k}'\cdot\mathbf{r}})\cdot(-\frac{q^2}{4\pi\epsilon_s r}e^{-\frac{r}{L_D}})\cdot(\frac{1}{\sqrt{V}}e^{i\mathbf{k}\cdot\mathbf{r}}) \\ &= \frac{q^2}{\epsilon_s V}\frac{1}{\frac{1}{L_D^2} + |\mathbf{k}-\mathbf{k}'|^2}, \end{aligned} \tag{16.32}$$

---

[15] Such situations arise when there are transitions between bands, say from the conduction to the valence band in interband transitions involving tunneling or optical optical processes, or in avalanche multiplication at high electric fields. The periodic part of the Bloch functions must also be included in intervalley scattering processes when a large part of the Brillouin zone is sampled. This may happen in transitions from the Γ-point with $|s\rangle$ orbital lattice-periodic states, to a $|p\rangle$ state in the same band, but at the other extrema, say near the Brillouin zone edge.

[16] The form of the screened Coulomb potential is identical to the Yukawa potential that keeps protons from flying apart in a nucleus. Hideki Yukawa proposed this potential in 1935. The particles that mediate the nuclear interaction called mesons were discovered experimentally, earning him the Nobel Prize in 1949.

where the integral is evaluated in 3D spherical coordinates over the entire space by aligning the vector $\mathbf{k}-\mathbf{k}'$ at an angle θ to the z-axis. Note that the scattering matrix element between plane-wave like envelope states is equal to the Fourier transform of the scattering potential. Now the total momentum scattering rate of state $\mathbf{k}$ is

$$\frac{1}{\tau_m(\mathbf{k})} = \sum_{\mathbf{k}'} \frac{2\pi}{\hbar} \cdot \Big|\frac{q^2}{\epsilon_s V} \frac{1}{\frac{1}{L_D^2} + |\mathbf{k}-\mathbf{k}'|^2}\Big|^2 \cdot \delta[E(\mathbf{k}') - (E(\mathbf{k}) \pm \hbar\omega)] \cdot (1-\cos\theta), \tag{16.33}$$

which upon summing over all 3D $\mathbf{k}'$ states, using the Dirac delta function, and the considering the effect of a total of N_{imp} impurities results in

$$\frac{1}{\tau_m(\mathbf{k})} = \frac{g_s g_v q^4 m_c^\star}{8\pi\epsilon_s^2 V} \cdot \underbrace{\frac{N_{imp}}{V}}_{n_{imp}} \cdot \frac{1}{(\hbar k)^3} \cdot [\ln(1+4k^2L_D^2) - \frac{4k^2L_D^2}{1+4k^2L_D^2}], \tag{16.34}$$

where $k = |\mathbf{k}|$ is the length of the wavevector, and $n_{imp} = N_{imp}/V$ is the volume density of the uncorrelated scatterers. The dimensions may be checked to be 1/sec. The term in the square brackets is a slowly varying function. The momentum scattering rate can now be used in Equation 16.29 to get the electron mobility explicitly. For a parabolic bandstructure $E(k) = \frac{\hbar^2k^2}{2m_c^\star}$ assuming a reference $E_c = 0$, the momentum scattering rate as a function of the electron kinetic energy is

$$\frac{1}{\tau_m(E)} = \frac{g_s g_v q^4 m_c^\star}{8\pi\epsilon_s^2} \cdot n_{imp} \cdot \frac{1}{(2m_c^\star E)^{\frac{3}{2}}} \cdot [\ln(1+\frac{8m_c^\star E}{\hbar^2}L_D^2) - \frac{\frac{8m_c^\star E}{\hbar^2}L_D^2}{1+\frac{8m_c^\star E}{\hbar^2}L_D^2}], \tag{16.35}$$

which indicates that the scattering rate decreases as the kinetic energy of electrons increases, and increases when the impurity density increases. These trends are indicated in Fig. 16.12. For a non-degenerately doped semiconductor, using Maxwell–Boltzmann approximation of the equilibrium distribution function $f_{\mathbf{k}}^0 \approx e^{-\frac{E(\mathbf{k})}{k_bT}}$, the electron mobility due to ionized impurity scattering is obtained to be

$$\begin{aligned} \mu_{imp} &\approx \frac{2^{\frac{7}{2}}(4\pi\epsilon_s)^2(k_bT)^{\frac{3}{2}}}{\pi^{\frac{3}{2}}q^3\sqrt{m_c^\star}n_{imp}F(\beta)} \sim \frac{T^{\frac{3}{2}}}{n_{imp}} \\ \mu_{imp} &\approx \frac{383082}{F(\beta)} \cdot \frac{(\frac{\epsilon_r}{10})^2}{\sqrt{\frac{m^\star}{0.2m_e}}} \cdot \frac{(\frac{T}{300\text{ K}})^{3/2}}{(\frac{n_{imp}}{10^{16}\text{cm}^3})} \cdot \frac{\text{cm}^2}{\text{V}\cdot\text{s}}, \end{aligned} \tag{16.36}$$

where $F(\beta) = \ln[1+\beta^2] - \frac{\beta^2}{1+\beta^2}$, and $\beta = 2\sqrt{\frac{2m_c^\star(3k_bT)}{\hbar^2}}L_D \approx 48.2 \times \sqrt{\frac{\epsilon_r}{10}} \cdot (\frac{T}{300\text{ K}}) \cdot \sqrt{\frac{m^\star}{0.2m_e}} \cdot \sqrt{\frac{10^{16}\text{cm}^{-3}}{n}}$ are dimensionless parameters[17]. As the impurity density increases, the mobility decreases due to more frequent scattering. The mobility increases with temperature because an

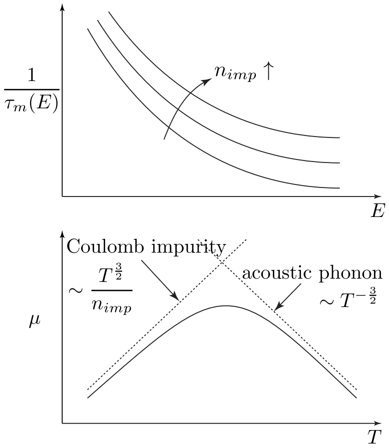


Fig. 16.12 Momentum scattering rate due to ionized impurities as a function of electron energy, and electron mobility as a function of temperature due to ionized impurity scattering and acoustic phonon scattering in three dimensions.

---

[17] The mobility due to ionized impurity scattering can be read off from the dimensionless form of Equation 16.36. For example for a semiconductor of conduction band effective mass $m_c^\star = 0.2m_e$, $\epsilon_r = 10$, and impurity density = free–electron concentration ($n_{imp} = n = 10^{17}/\text{cm}^3$, meaning the semiconductor is uncompensated), $\beta = 83.5/\sqrt{10} = 26.4$, and $F(\beta) = 5.55$. The mobility at room temperature then is $\mu_{imp} \approx \frac{383082}{5.55} \cdot 1 \cdot \frac{1}{10} \approx 6902$ cm^2/V·s. This is the mobility if only ionized impurity scattering was active, meaning phonon and other scattering mechanisms are not considered. Similar expressions for other scattering mechanisms will be discussed in Chapter 23.

energetic electron is less perturbed by the scattering potential from its original path. The mobility limited by ionized impurity scattering is high for semiconductors with small effective masses, and for large dielectric constants. For non-degenerate electron concentrations, the ionized impurity scattering limited mobility increases with temperature as $T^{+3/2}$ as indicated in Fig. 16.12. When the scattering rate and mobility limited by acoustic phonon scattering is calculated in a similar manner, the dependence is found to be $\mu_{ac} \sim T^{-3/2}$. Since all scattering mechanisms act in parallel, the rates add, and the net mobility is due to the dominant scatterers. At low temperatures, phonons are frozen out, and defects and impurity scattering dominate. At high temperatures in sufficiently defect free samples, phonon scattering dominates. We will evaluate the scattering rates and mobilities of various processes in more detail in **Chapter 23**.

16.9 Semiconductor electron energies for photonics

In the remainder of this chapter, we selectively sample the contents of **Module IV**, **Chapters 26-29** on quantum photonics with semiconductors. A photon interacts with a semiconductor for the most part in an identical fashion as it does with an atom. Because the semiconductor is made of a large number of atoms, the primary difference is a larger number of allowed transitions compared to the atom. There are a few new physical phenomena that emerge in a semiconductor crystal and cannot happen for an individual atom, but knowing the photon-atom interaction physics goes a long way in explaining the photonic properties of semiconductors. Fig. 16.13 and its caption discusses the interaction of photons with atoms and with semiconductors in a comparative fashion.

Let the semiconductor bandgap be $E_c - E_v = E_g$, where E_c is the conduction band minimum, and E_v is the valence band maximum. Let the conduction band effective mass be $m_c^\star$ and the valence band effective mass $m_v^\star$, so that the conduction band states are given by $E_c(k) = E_c + \frac{\hbar^2 k_c^2}{2m_c^\star}$ and the valence band states are $E_v(k) = E_v - \frac{\hbar^2 k_v^2}{2m_v^\star}$. This is indicated in Fig. 16.14. Now if there is an optical transition between the conduction band state E_2 and a valence band state E_1, the photon energy must be $h\nu = E_2 - E_1$ to ensure energy conservation. Because the photon in circularly polarized light carries an angular momentum $\pm\hbar$, the conduction band and the valence band should have a net angular momentum difference of $\hbar$, which they do – because the conduction band states derive from $|s\rangle$ orbitals, and the valence band states from $|p\rangle$ orbitals. Finally, to ensure momentum conservation, we must have $\hbar\mathbf{k}_c = \hbar\mathbf{k}_v + \hbar\mathbf{k}_p$, where $\hbar\mathbf{k}_p$ is the photon momentum. Since the electron states have $|\mathbf{k}_c| \approx |\mathbf{k}_v| \approx \frac{\pi}{a_0}$ where a_0 is a lattice constant, whereas the photon momentum $\hbar|\mathbf{k}_p| = \frac{h}{\lambda}$ is *much* smaller.

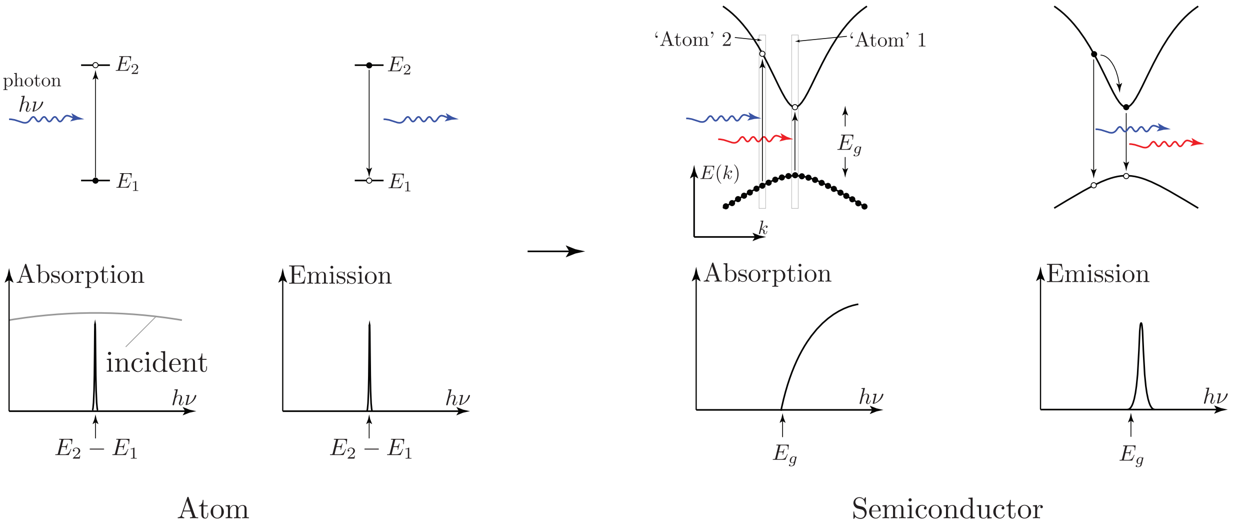


Fig. 16.13 Since an atom allows sharp and discrete energy levels for electrons (say E_1 and E_2 as shown), if photons of a large spectrum of energies $\hbar\omega = h\nu$ are incident, only those photons that sharply match the energy conservation requirement $E_2 - E_1 = h\nu$ are absorbed by the atom. An electron must occupy state E_1 and state E_2 must be empty for the absorption to occur. After absorption, the atom is in an excited state – it takes typically a nanosecond for the electron to spontaneously transition back from E_2 to E_1, emitting the photon back. This process of spontaneous emission can be slowed down by placing the atom in an optical cavity of an appropriate size. The emitted spectrum is also sharp. A single atom can only emit one photon at a time. Compared to the atom, as we have discussed in earlier chapters, a direct-gap semiconductor crystal has many energies in bands, which derive from the atomic states. In an oversimplified picture, the eigenvalues of the bands at each k may be pictured as separate atomic states. The absorption is vertical in k. Since the density of states increases away from the band edges, more photons at each energy can be absorbed, leading to an absorption edge at the bandgap E_g, and increasing absorption for higher photon energies. The absorption spectrum depends on the semiconductor material as well as dimensionality: it is different for 3D, 2D, 1D, and 0D. After absorption, the semiconductor is in an excited state. An electron in a high energy state in the conduction band can directly recombine with the hole state in the valence band of the same k by emitting a photon, or it can emit phonons and relax down in the conduction band to lower energy states. Since phonon emission takes ∼ps whereas photon emission takes ∼ns, most electrons relax to near the bottom of the conduction band, and most holes relax to the top of the valence band. They then recombine to emit photons. The emission spectrum is therefore peaked near the bandgap, quite different from the absorption spectrum. The linewidth of the emission spectrum is also larger than the atomic states because of the large number of states available close in energy. A bulk semiconductor can emit several photons of the same energy at the same time because there are a large number of states that simultaneously match the energy and momentum criteria.

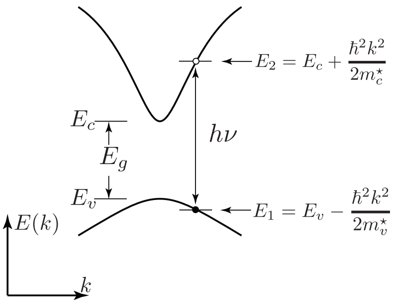

$E_2 = E_c + \frac{\hbar^2 k^2}{2m_c^\star}$
E_c
E_g
$h\nu$
E_v
$E(k)$
$E_1 = E_v - \frac{\hbar^2 k^2}{2m_v^\star}$
k

Fig. 16.14 Electron energies in bands for which optical transitions are allowed.

While this approximation is a very good one for IR, visible, and UV photons, it fails if the photon energy is too high. For example, a X-ray photon can possess a substantial momentum comparable or even larger than $\hbar/a_0$ where a_0 is a lattice constant, and since such a photon also has a very large energy, it can excite core-level electrons into vacuum in addition to the electrons in bands. This is the principle of X-ray photoelectron spectroscopy or XPS measurements. Considering photons of smaller energies, we make the assumption $k_c = k_v = k$, i.e., **the conduction and valence band states that talk to, i.e., absorb or emit photons must have the same** k. Pictorially, this is what we call "vertical" transitions in the $E(k) - k$ diagram. In Fig. 16.14, the energy state E_2 in the conduction band and one in the valence band E_1

of the same k are

$$E_2 = E_c + \frac{\hbar^2 k^2}{2m_c^\star}, \text{ and } E_1 = E_v - \frac{\hbar^2 k^2}{2m_v^\star}, \tag{16.37}$$

which leads to a photon energy

$$h\nu = E_2 - E_1 = (E_c - E_v) + \frac{\hbar^2}{2}(\frac{1}{m_c^\star} + \frac{1}{m_v^\star})k^2 = E_g + \frac{\hbar^2 k^2}{2m_r^\star}, \tag{16.38}$$

where $m_r^\star = m_c^\star m_v^\star / (m_c^\star + m_v^\star)$ is a reduced effective mass [18]. Thus,

$$\underbrace{h\nu}_{\text{light}} = \underbrace{E_g + \frac{\hbar^2 k^2}{2m_r^\star}}_{\text{matter}} \implies \boxed{\frac{\hbar^2 k^2}{2m_r^\star} = h\nu - E_g.} \tag{16.39}$$

Because the LHS of the boxed equation is positive, the above calculation reinforces our intuitive observation that only photons of energy $h\nu$ equal or larger than the energy bandgap E_g will interact with the electron states[19]. Unlike the atom, in a semiconductor there are a large number of transitions possible for various k values that produce a photon of the same energy $h\nu$, which we quantify in the next section.

16.10 The optical joint density of states $\rho_J(\nu)$

How many such electron k states are available per unit volume of the semiconductor to emit, or absorb photons in the energy interval[20] $[h\nu, h(\nu + d\nu)]$? This quantity is of central importance in the evaluation of electron-photon interactions, and is called the **optical joint density of states** (JDOS) of the semiconductor, denoted by $\rho_J(\nu)$. The number of states per unit volume is $\rho_J(\nu)d(h\nu)$. The optical JDOS $\rho_J(\nu)$ captures the property of light (through $h\nu$), and matter (through the semiconductor parameters the bandgap E_g, and band effective masses $m_c^\star$ and $m_v^\star$). The quantity $\rho_J(\nu)$ will also reflect if the optically active semiconductor electron states are unconfined and free to move in 3D bulk, or confined in quantum wells (2D), quantum-wires (1D), or quantum dots (0D)[21].

To count the number of electron states in 3D, assume the electrons are confined in a cubic box of size $L_x = L_y = L_z = L$ as shown in Fig. 16.15. Because an integer number of electron half-wavelengths must fit in the box, we get the condition $\mathbf{k} = (k_x, k_y, k_z) = \frac{\pi}{L}(n_x, n_y, n_z)$, where (n_x, n_y, n_z) is a triplet of integers each of whose values can be $0, 1, 2, ...$. This defines a 3D lattice in the $\mathbf{k} = (k_x, k_y, k_z)$ space in the first octant, where each point denotes an allowed electron state, occupying a volume $(\frac{\pi}{L})^3 = \frac{\pi^3}{V}$. Because the volume of the semiconductor cube is $V = L^3$ and we are interested in the JDOS $\rho_J(\nu)$ *per unit volume*, we also define the *total* JDOS as $D_J(\nu)$. The JDOS per unit volume is then related to the total JDOS by $\rho_J(\nu) = D_J(\nu)/V$.

---

[18] The reduced effective mass $m_r^\star$ accounts for the fact that both the valence and conduction bands participate in an interband transition

[19] We are neglecting excitonic, multiparticle, and tunneling effects here and focusing on the strongest interband transitions.

[20] We use $\hbar\omega = h\nu$ interchangeably.

[21] Quantum confined semiconductor structures are routinely used in practical semiconductor light emitting diodes (LEDs), semiconductor optical amplifiers (SOAs), optical modulators, and lasers that are discussed in Chapters 28 and 29.

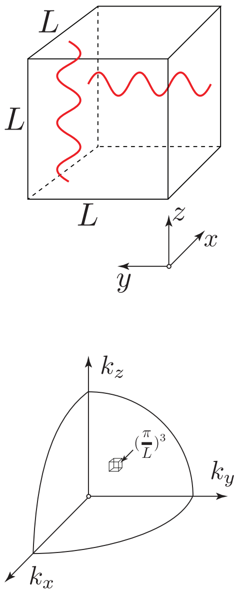


Fig. 16.15 Electron states in a cube of side L that can interact with photons.

We discuss the 3D optical JDOS first. The optical JDOS $\rho_\nu(k)$ in the **k**-space counts the same states as in the energy space, meaning $\rho_k(k)dk = D_J(\nu)d(h\nu)$. Using Equation 16.39 in the form $\frac{\hbar^2}{2m_r^\star}(k_x^2 + k_y^2 + k_z^2) = h\nu - E_g$, we obtain

$$\rho_k^{3D}(k)dk = 2 \cdot \frac{1}{8} \cdot \frac{4\pi k^2 dk}{\frac{\pi}{L_x}\frac{\pi}{L_y}\frac{\pi}{L_z}} = D_J(\nu)d(h\nu)$$

$$\implies \boxed{\rho_J^{3D}(\nu) = \frac{1}{2\pi^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{h\nu - E_g},} \qquad (16.40)$$

where $\rho_J^{3D}(\nu) = D_J(\nu)/V$ was used. The JDOS $\rho_J^{3D}(\nu)$ has units of $1/(\text{energy} \cdot \text{Volume})$ as may be verified from Equation 27.44. For semiconductors it is expressed in $\frac{1}{\text{eV}\cdot\text{cm}^3}$. The JDOS for quantum confined electron states is found in the same manner as for the 3D, except instead of a 3D k space, we have a 2D, or lower-dimensional k space. For example, for a quantum well laser, the electron and hole states are free to move in two dimensions (x, y) because of large dimensions $L_x = L_y = L$ and confined in the third z direction because of a heterostructure quantum well confining potential to a length L_z smaller than the electron de-Broglie wavelengths, leading to the requirement $k_z = n_z \frac{\pi}{L_z}$. This situation is realized if we create a structure in which $L_z << L$, where L_z is a few nm, whereas L is the size of the wafer – in a few 100s of μm or in mm (see Fig. 16.16).

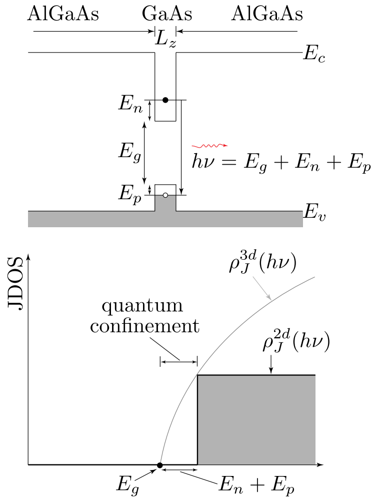


Fig. 16.16 A semiconductor heterostructure quantum well and its optical joint DOS.

For such a quantum-confined structure, Equation 16.39 now takes the form

$$\frac{\hbar^2}{2m_r^\star}(\underbrace{k_x^2 + k_y^2}_{k_\parallel^2} + k_z^2) = h\nu - E_g$$

$$\implies \frac{\hbar^2}{2m_r^\star}k_\parallel^2 = h\nu - [E_g + \underbrace{\frac{\hbar^2}{2m_e^\star}(\frac{\pi}{L_z}n_z)^2}_{E_n} + \underbrace{\frac{\hbar^2}{2m_h^\star}(\frac{\pi}{L_z}n_z)^2}_{E_p}], \qquad (16.41)$$

where we identify the blue-shift in the minimum transition energy due to quantum confinement in the conduction and the valence bands $E_n + E_p$. This quantum-engineering of the effective bandgap is a very powerful tool for engineering light-matter interactions in semiconductor heterostructures[22].

The allowed electron modes then form a dense 2D lattice in the $(k_x, k_y) = \frac{\pi}{L}(n_x, n_y)$ planes, separated by large $k_z = \frac{\pi}{L_z}n_z$ in the k_z direction. Then, the "cylindrical" four-fold symmetry around the k_z-axis of the allowed lattice of mode points in **k**-space has each mode occupying a volume $\frac{\pi}{L_x}\frac{\pi}{L_y}\frac{\pi}{L_z}$, and the JDOS is given by

$$\rho_k^{2D}(k)dk = 2\cdot\frac{1}{4}\cdot\frac{(2\pi k_{\|}dk_{\|})\frac{\pi}{L_z}}{\frac{\pi}{L_x}\frac{\pi}{L_y}\frac{\pi}{L_z}} = D_J(\nu)d(h\nu)$$

$$\implies \boxed{\rho_J^{2D}(\nu) = \frac{m_r^\star}{\pi\hbar^2 L_z}\Theta[h\nu-(E_g+E_n+E_p)],} \tag{16.42}$$

where we again used $\rho_J^{2D}(\nu) = D_J(\nu)/V$. The Heaviside unit-step function $\Theta[...]$ in the 2D optical JDOS yet again reinforces our intuition that the smallest photon energy that can interact with the quantum well electron states must have an energy $h\nu = E_g + E_n + E_p$ because quantum confinement has pushed the minimum energy in the conduction band up to $E_c + E_n$ and in the valence band down to $E_v - E_p$. The units of $\rho_J^{2D}(\nu)$ is still $\frac{1}{\mathrm{eV\cdot cm^3}}$, the same as for 3D – this is verified from Equation 16.42.

Continuing to lower dimensions, it is left to the reader to show that for electron states in 1D quantum wires confined in the (x, y) plane to (L_x, L_y) and free to move in the z-direction, the optical JDOS is

$$\boxed{\rho_J^{1D}(\nu) = \frac{2}{\pi L_x L_y}\sqrt{\frac{2m_r^\star}{\hbar^2}}\frac{1}{\sqrt{h\nu-(E_g+E_n+E_p)}},} \tag{16.43}$$

where the quantum confinement energies are now $E_n = \frac{\hbar^2}{2m_e^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2]$ and $E_p = \frac{\hbar^2}{2m_h^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2]$.

For 0D quantum dots in which electrons are confined in all three directions to boxes of size (L_x, L_y, L_z), the structure mimics an atomic system, giving an optical JDOS

$$\boxed{\rho_J^{0D}(\nu) = \frac{2}{L_x L_y L_z}\delta[h\nu-(E_g+E_n+E_p)],} \tag{16.44}$$

where $\delta[...]$ is the Dirac delta function highlighting that the electron energies are discrete and atomic-like, and do not form extended states or bands. The quantum confinement energies are now $E_n = \frac{\hbar^2}{2m_e^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2 + (\frac{\pi}{L_z}n_z)^2]$ and $E_p = \frac{\hbar^2}{2m_h^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2 + (\frac{\pi}{L_z}n_z)^2]$. It may be verified again that the units of $\rho_J^{0D}(\nu)$ are exactly the same as in other dimensions. Fig. 16.17 shows the JDOS of all dimensions.

We will shortly see that the optical gain coefficient of a semiconductor in 1/cm units is given by $g_0(\nu) = A\cdot[\frac{\lambda_0^2}{8\pi n^2}]\cdot[h\rho_J(\nu)]\cdot[f_c(E_2) - f_v(E_1)]$, where A is the Einstein coefficient of units 1/sec, λ_0 is the free-space photon wavelength, and n is the refractive index of the semiconductor. **The optical gain coefficient is characteristically proportional to the optical joint density of states** $\rho_J(\nu)$. Those conversant with atomic laser theory will immediately recognize the similarity to the

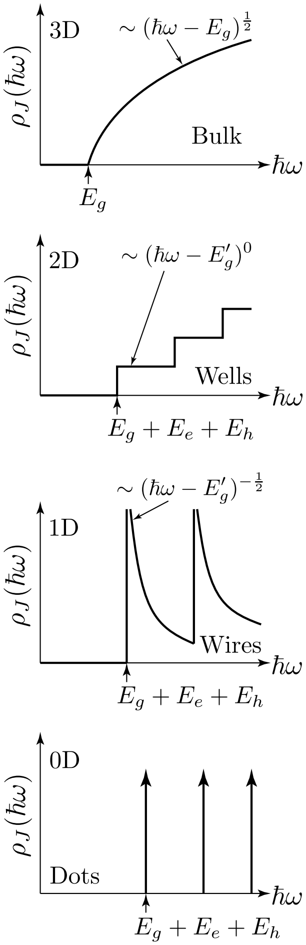


Fig. 16.17 Optical joint density of states for 3D to 0D semiconductor structures for parabolic conduction and valence bands. Note that the absorption coefficient is related to the JDOS via Equation 27.43. This Fig. is repeated and discussed in detail in Chapter 27.

gain coefficient $g_0(\nu) = A \cdot [\frac{\lambda_0^2}{8\pi n^2}] \cdot [\mathcal{L}(\nu)] \cdot [N_2 - N_1]$, where $\mathcal{L}(\nu)$ is the **lineshape** function in units of time (seconds), and $N_2 - N_1$ is the population inversion in $1/\text{cm}^3$ units. The semiconductor counterpart of the lineshape function $h\rho_J(\nu)$ has units sec/cm^3, and the population inversion term motivates a discussion of the dimensionless occupation probabilities $f_c(E_2)$ and $f_v(E_1)$ of the allowed conduction and valence band electron energy states, to which we now turn.

16.11 Occupation of electron states for photonics

Because electrons are fermions, the Pauli exclusion principle forces them to follow the Fermi–Dirac distribution function in energy. If the semiconductor is in thermal equilibrium at a temperature T, the distribution of electrons in the semiconductor in energy is given by

$$f(E) = \frac{1}{1 + e^{\frac{E - E_F}{k_b T}}}, \tag{16.45}$$

where the Fermi energy E_F is a single energy characterizing the number of electrons, and k_b is the Boltzmann constant. However, for semiconductor photonic devices, we will be interested in pn junction diodes under electrical bias conditions. In a pn junction, we will typically have the following situation: the conduction band electrons in the n-side of the junction will be connected to an ohmic metal contact lead, and the valence band electrons in the p-side of the diode to another ohmic metal contact lead. These two leads will then be connected to a battery of voltage V. Application of a voltage V pulls the diode out of the state of equilibrium, as shown in Fig. 16.18. We can now consider two *separate* electron distributions: those of electrons in the conduction band, and of electrons in the valence band. These two electron systems are in equilibrium amongst themselves, but they are out of equilibrium from each other. The electrons in the conduction band in the n-side of the diode are in equilibrium with the voltage of the battery terminal connected to them; this is modeled by assigning to them a *quasi Fermi level* F_n. Similarly the electron states in the valence band of the p-side of the diode are in equilibrium with the other terminal of the battery, sharing a quasi Fermi level F_p. The electron populations at the two terminals of the battery are held at an energy qV apart, meaning $F_n - F_p = qV$.

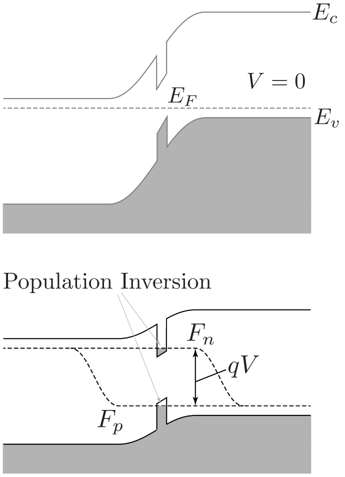


Fig. 16.18 A pn junction diode with a quantum well inside the depletion region. Such a structure is at the heart of semiconductor LEDs and lasers. When a voltage V is applied, the Fermi level splits into quasi Fermi levels of electrons F_n and holes F_p. If the quasi Fermi levels enter the bands, then population inversion is achieved, because there are electrons in the conduction band and holes in the valence bands in a region of the semiconductor where the occupation should be opposite under equilibrium.

Because of the above definitions, the occupation probability of an electron state of energy E_2 in the conduction band is $f_c(E_2)$, and the occupation probability of an electron state of energy E_1 in the valence band is $f_v(E_1)$, where each is a Fermi–Dirac function modified to account for the non-equilibrium physics:

$$\boxed{f_c(E_2) = \frac{1}{1+e^{\frac{E_2-F_n}{k_bT}}}} \text{ and } \boxed{f_v(E_1) = \frac{1}{1+e^{\frac{E_1-F_p}{k_bT}}}.} \tag{16.46}$$

If we are interested in the 3D electron density in the conduction band, we can use the 3D conduction band density of states to write

$$n = \int_{E_c}^{\infty} dE \cdot \rho_c(E) f_c(E) = \int_{E_c}^{\infty} dE \cdot \frac{1}{2\pi^2}(\frac{2m_c}{\hbar^2})^{\frac{3}{2}}\sqrt{E-E_c}\frac{1}{1+e^{\frac{E-F_n}{k_bT}}},$$
$$n_d = N_c^d F_{\frac{d-2}{2}}(\frac{F_n - E_c}{k_bT}) \text{ and } p_d = N_v^d F_{\frac{d-2}{2}}(\frac{E_v - F_p}{k_bT}). \text{ for d dim.} \tag{16.47}$$

For finding the probability of E_1 being an empty energy state in the valence band (or hole), we must use the probability that the electron state is *not occupied*, which is precisely $1 - f_v(E_1)$. Let us say that for photon emission, we need an electron occupying energy E_2 in the conduction band to transition into an empty state E_1 in the valence band such that $E_2 - E_1 = h\nu$. The probability for the electron states to satisfy the criteria is the product $f_c(E_2)[1 - f_v(E_1)]$.

16.12 Absorption, and emission: spontaneous and stimulated

Let us assume that we have identified the joint density of states $\rho_J(\nu)$ of a material: for example, a semiconductor. Einstein re-derived Planck blackbody radiation formula by introducing the concepts of rates of spontaneous emission, stimulated emission, and absorption through the net rate equation for *any* material:

$$\frac{1}{d(h\nu)}\frac{d\Delta n}{dt} = -\underbrace{A\rho_J(\nu)f_c(E_2)[1-f_v(E_1)]}_{=R_{sp}(\nu),\text{ Spontaneous Emission}}$$
$$-\underbrace{B\rho_J(\nu)\rho_\nu f_c(E_2)[1-f_v(E_1)]}_{=R_{st}(\nu),\text{ Stimulated Emission}} + \underbrace{B\rho_J(\nu)\rho_\nu f_v(E_1)[1-f_c(E_2)]}_{=R_{ab}(\nu),\text{ Absorption}}, \tag{16.48}$$

where the LHS tracks the rate of change of excess electron (or hole) density Δn per energy interval $d(h\nu)$. Electrons may be lost from this level due to spontaneous emission of photons of energy $h\nu = E_2 - E_1$ if the electron transitions to empty state E_1 in the valence band, as indicated in Fig. 16.19. This will happen at a rate $R_{sp}(\nu)$, as indicated in Equation 16.48. The net spontaneous emission rate $R_{sp}(\nu)$ is made of the JDOS $\rho_J(\nu)$, which counts how many electron state pairs (E_2, E_1) are available per unit energy interval per unit volume satisfying $h\nu = E_2 - E_1$, and the Fermi factor $f_c(E_2)[1 - f_v(E_1)]$, which is the

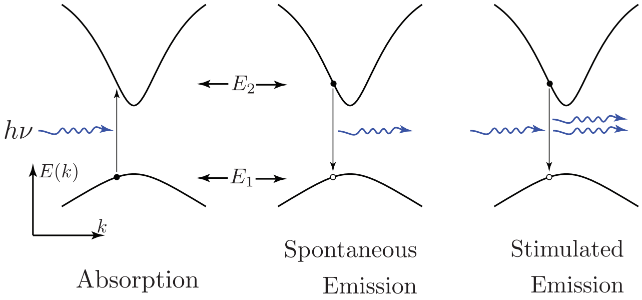


Fig. 16.19 Absorption, and spontaneous and stimulated emission processes in a semiconductor. While the rates of absorption and stimulated emission must be proportional to the number of photons already present, that of spontaneous emission should not depend on it. This fact is highlighted in Equation 16.48.

probability that the occupations of the states are just right for emitting a photon. The coefficient A, an effective spontaneous emission rate, at this stage may be considered something that can be obtained from experiments. The exact form of A may be calculated from time-dependent perturbation theory, and more accurately from quantum electrodynamics, to be discussed in **Chapter 29**. It is a rate constant typically of the order $1/\mathrm{ns}$ for most semiconductors.

The second term in the RHS of Equation 16.48 is the rate of *stimulated emission*, $R_{st}(\nu)$. The coefficient is B – which has different units from the A coefficient. Einstein showed that for the above picture of light-matter interaction to be consistent with Planck's blackbody radiation formula, the A and B coefficients are related by

$$\boxed{\frac{A}{B} = \frac{8\pi n^3 \nu^3 h}{c^3}.} \tag{16.49}$$

The JDOS and probability functions are identical for the stimulated emission and spontaneous emission rates. Other than the A and B coefficients, the two processes are distinguished by the term ρ_ν, which is the energy density of the *photon field*, in units of $\mathrm{eV/cm^3}$. It is related to the *intensity* of the photon field I_ν by $I_\nu = \rho_\nu \cdot (c/n)$, where c/n is the photon velocity in a medium of refractive index n. The intensity of the photon field is obtained from the Poynting vector $\mathbf{S}_\nu = \mathbf{E}_\nu \times \mathbf{H}_\nu$ from the electric and magnetic fields, the magnitude of which is $I_\nu = \frac{E_\nu^2}{2\eta}$ in units of $\mathrm{Watts/cm^2}$, or the rate of flow of energy per unit area. Here E_ν is the electric field amplitude, and η the electromagnetic wave impedance of mode ν.

The spontaneous emission rate $R_{sp}(\nu)$ does not require photons to be present to induce emission, true to its name. The spontaneously emitted photon goes into any of the allowed photon modes in the electromagnetic field: this is a random process and there is no correlation of which mode one spontaneously emitted photon goes with the next. The stimulated emission rate $R_{st}(\nu)$ and the absorption rate $R_{ab}(\nu)$ are both proportional to the photon field ρ_ν implying they cannot occur in the absence of light. **The photon created by stimulated emission goes into exactly the same electromagnetic mode as the photon that stimulated it**. The emitted photon has the same phase, wavelength, polarization, and is coherent with the photon that stimulated it: **it is a clone**! The process of stimulated emission is responsible for gain and lasing. In stimulated emission, we start with one photon but end with two – implying *gain*.

Typically under optical amplification and lasing conditions, the spontaneous emission is a low-intensity background. Let us look specifically at the stimulated emission and absorption processes. The net emission rate or downward transitions is given by:

$$R_{em}(\nu) = \underbrace{R_{st}(\nu)}_{\downarrow} - \underbrace{R_{ab}(\nu)}_{\uparrow}$$
$$= B\rho_J(\nu)\rho_\nu(f_c(E_2)[1-f_c(E_1)] - f_v(E_1)[1-f_c(E_1)])$$
$$= B\rho_J(\nu)\rho_\nu[f_c(E_2) - f_v(E_1)]. \qquad (16.50)$$

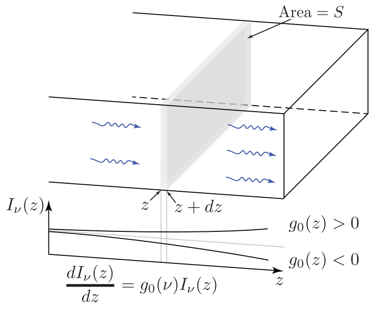


Fig. 16.20 The photon (or light) density $I_\nu(z)$ changes in a semiconductor as the wave propagates. If $g_0(z) < 0$, there is photon loss by absorption, which is central to the operation of photodetectors and solar cells. If $g_0(z) > 0$, there is optical gain. Optical gain is possible by stimulated emission, and is central to the operation of semiconductor optical amplifiers, and lasers.

Because each such stimulated emission event increases the photon count by one, and the energy of the photon field by $h\nu$, we should track the energy flowing though a box of cross-section area S, as shown in Fig. 16.20. The increase in the photon number in a volume $S \cdot dz$ results in an increase in the photon field intensity

$$[I_\nu(z+dz) - I_\nu(z)] \cdot S = R_{em}(\nu) \cdot h\nu \cdot S \cdot dz \implies \frac{dI_\nu(z)}{dz} = R_{em}(\nu)h\nu, \qquad (16.51)$$

which with the relation between the Einstein A and B coefficients, and the relation $I_\nu(z) = \rho_\nu \cdot c$ leads to

$$\frac{dI_\nu(z)}{dz} = \underbrace{A \cdot \frac{\lambda_0^2}{8\pi n^2} \cdot h\rho_J(\nu) \cdot [f_c(E_2) - f_v(E_1)]}_{g_0(\nu)} \cdot I_\nu(z) = g_0(\nu)I_\nu(z), \qquad (16.52)$$

which indicates the semiconductor gain coefficient is given by

$$\boxed{g_0(\nu) = \underbrace{A \cdot \frac{\lambda_0^2}{8\pi n^2} \cdot h\rho_J(\nu)}_{\alpha_0(\nu),\text{ depends on material \& } h\nu} \cdot \underbrace{[f_c(E_2) - f_v(E_1)]}_{\text{population inversion}}.} \qquad (16.53)$$

If $g_0(\nu) > 0$, the intensity $I_\nu(z)$ grows with distance according to $I_\nu(z) = I_\nu(0)e^{g_0(\nu)z}$ and we have photon gain, as indicated in Fig. 16.20. If $g_0(\nu) < 0$, photons are absorbed, and one should refer to this quantity as the absorption coefficient. The prefactor $\alpha_0(\nu)$ in Equation 16.53 for the gain coefficient that depends on the semiconductor JDOS and the photon properties is always positive. So in order to obtain gain, the Fermi difference function should meet the criteria

$$f_c(E_2) - f_v(E_1) = \frac{1}{1+e^{\frac{E_2-F_n}{k_bT}}} - \frac{1}{1+e^{\frac{E_1-F_p}{k_bT}}} > 0$$
$$\implies F_n - F_p > E_2 - E_1 \implies \boxed{F_n - F_p > h\nu > E_g.} \qquad (16.54)$$

The boxed condition $F_n - F_p > h\nu > E_g$ is the **population inversion** criteria for semiconductors, or the condition to achieve optical gain. It is also referred to as the **Bernard–Duraffourg** condition after those who first identified it. Fig. 16.21 indicates the semiconductor gain coefficient at two excitation conditions. Since in a pn diode $F_n - F_p = qV$, it is necessary to apply a voltage larger than the effective bandgap of the optically active region to obtain gain. For photon energies matching $F_n - F_p = h\nu$, the net gain coefficient $g_0(\nu) = 0$, meaning the semiconductor medium is transparent to those photons, as seen in Fig. 16.21.

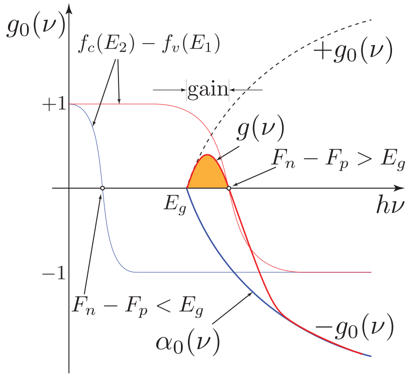


Fig. 16.21 The gain spectrum of a bulk semiconductor at non-equilibrium conditions. When the quasi-Fermi level $F_n - F_p$ splitting is smaller than the bandgap, the gain spectrum is simply $g_0(\nu) = -\alpha_0(\nu)$, the absorption coefficient. When $F_n - F_p > h\nu > E_g$, population inversion is achieved and the gain spectrum is +ve over a window of photon energies $E_g < h\nu < F_n - F_p$. For higher photon energies, there is absorption and loss.

Note that this is an **induced transparency**, meaning if the semiconductor had been in equilibrium, it would normally absorb photons of energy $h\nu = F_n - F_p > E_g$ because it is larger than the bandgap. For photon energies exceeding this quasi-Fermi level split, the Fermi difference function is negative, leading to loss. Thus, the semiconductor gain spectrum has a bandwidth $\frac{E_g}{h} \leq \nu \leq \frac{F_n - F_p}{h}$, and the shape $g_0(\nu)$ dictated by the product of the optical JDOS, and the Fermi difference function. The Fermi difference functions are shown for a low-level carrier injection into the pn junction when $F_n - F_p < E_g$, and a high level injection when $F_n - F_p > E_g$, together with the corresponding gain spectra. If photon energies that fall in the gain window are reflected back and forth between mirrors, the gain builds up and exceeds the loss, resulting in stimulated emission. This is the working principle of the **semiconductor laser**, the topic of **Chapter 29**. We will discuss these photonic properties of bulk and low-dimensional semiconductor quantum structures in quantitative detail in the chapters of **Module IV**.

It is clear that at equilibrium, or for small levels of carrier injection when $F_n \approx F_p = E_F$, the factor $f_c(E_2) - f_v(E_1) \approx -1$, and we obtain the optical absorption coefficient of the semiconductor $\alpha_0(\nu) = A \cdot \frac{\lambda_0^2}{8\pi n^2} \cdot h\rho_J(\nu)$, whereby the intensity of photons changes as $\frac{dI_\nu(z)}{dz} = -\alpha_0(\nu) I_\nu(z)$ as they are absorbed. This leads to an intensity variation $I_\nu(z) = I_\nu(0) e^{-\alpha_0(\nu) z}$, implying that the intensity of light decreases exponentially with depth if the photons are absorbed. This decay happens over a characteristic length scale of $\alpha_0(z)$, which is defined as the absorption length[23].

Returning to spontaneous emission, we obtain a spectrum

$$R_{sp}(\nu) = A\rho_J(\nu) f_c(E_2)[1 - f_v(E_1)]$$

$$= A\rho_J(\nu) \frac{1}{1 + e^{\frac{E_2 - F_n}{k_b T}}} \cdot \frac{e^{\frac{E_1 - F_p}{k_b T}}}{1 + e^{\frac{E_1 - F_p}{k_b T}}} \approx A\rho_J(\nu) e^{-\frac{h\nu}{k_b T}} e^{\frac{F_n - F_p}{k_b T}}. \quad (16.55)$$

This condition is important for **LEDs**, because it indicates that by splitting the quasi-Fermi levels $F_n - F_p = qV$ with a voltage, we exponentially increase the spontaneous emission rate of photons. Of course these photons must be extracted before they get re-absorbed again, because though absorption rate is smaller, it is not negligible! Unlike the

[23] The exponential decay of light intensity in a material medium is called the classical **Beer–Lambert law**.

laser, the LED can operate at much smaller levels of current injection well below the conditions of population inversion. The photons emitted by a LED are due to spontaneous emission, which is why they lack the coherence and directivity of those from a laser. Incoherent light from LEDs finds as much use as coherent light from lasers.

16.13 Chapter summary section

In this chapter, we learned:

- The basic principles of semiconductor physics,
- Fundamental concepts for semiconductor electronic devices,
- Fundamental concepts for semiconductor photonic devices, and
- And are ready to put the physics to use in applications in electronic and photonic devices in **Modules III & IV**!

Further reading

Since this chapter serves as a condensed version of the entire subject, I have recommended several books of previous chapters. For Modules III and IV, I make a few more suggestions, which will be repeated in relevant chapters. For semiconductor devices, **Physics of Semiconductor Devices** by Simon Sze is the most comprehensive and remains as relevant today as when it was written. I strongly recommend **Wave Mechanics applied to Semiconductor Heterostructures** by Gerald Bastard, where the emphasis is less on devices, but more on the quantum physics. For those who want to delve deeper into the quantum mechanical foundations of semiconductors, I recommend **Quantum Theory of the Optical and Electronic Properties of Semiconductors** by Haug and Koch. They have taken a rigorous approach to the subject, and cover several optical phenomena that are not discussed in this book.

Exercises

(16.1) **Reading exercise**

Since this chapter is a summary of the past and future topics of the book, no exercises are assigned specific to this chapter. To take advantage of it acting as a bridge from the physics to device applications, a few reading exercises are assigned.

(a) William Shockley, a pioneer of semiconductor physics and the co-inventor of the transistor wrote a review article in 1952 called **Transistor Electronics: Imperfections, Unipolar and Analog Transistors**. This article captures the physics of semiconductors at a heuristic level that remains unmatched to date. Find and read this gem of an article. You will see the thought processes that underpin the subject require minimal mathematical complications to understand. You will also encounter how the names source, drain, and gate were introduced!

(b) An excellent (and *very* thick!) book on semiconductor devices is **Semiconductor Devices: Pioneering Papers**, a collection by Simon Sze. This collection is a treasure trove of papers dating back to the 1800s. Those who want to learn of the true history behind today's semiconductor devices will appreciate reading the early papers, and who knows, it may inspire you to invent something new!

Part III

Quantum Electronics with Semiconductors

Game of Modes: Quantized R, L, and C

17

17.1 Classical R, L, C circuits 376
17.2 Quantized conductance 377
17.3 Quantum capacitance 382
17.4 Kinetic inductance 388
17.5 Quantum R, L, C circuits 392
17.6 Negative R, L, and C 393
17.7 Chapter summary section 397
Further reading 397
Exercises 397

Traditional electronic circuit design is based on understanding and controlling the roles of resistors (R), inductors (L), and capacitors (C). Since the three RLC circuit elements are *passive*, they can *store* and *flow* electrical energy, but not indefinitely because they are unable to *amplify* electrical signals to counteract and recover the losses in the resistor. Recovery of the signal requires electronic *gain*. Historically, electronic gain was found in vacuum devices, which have been mostly miniaturized and replaced by semiconductor-based *active* devices. In the following chapters, we discuss the quantum physics of diodes and transistors, the primary semiconductor devices that provide electronic gain. In this chapter, we focus on the quantum physics of the passive elements. As long as the wavelength $\lambda = c/f$ of an electrical signal of frequency f (c is the speed of light) is much larger than the dimensions of a resistor, capacitor, or inductor, these passive elements behave as "point devices", or lumped elements of R, L, C. As electronic circuits are miniaturized and their operational frequencies increase to higher speeds, the quantum limits of resistance (or conductance), inductance and capacitance (Fig. 17.1) assume increasingly important roles. In this chapter, we learn the effect of the number of modes M, group velocities v, and density of states (DOS) on:

- Quantized conductance $G_q = \frac{q^2}{h} M$ and resistance $R_q = \frac{h}{Mq^2}$,
- Quantum capacitance $C_q = q^2 \cdot \mathrm{DOS} = \frac{q^2}{h} \langle \frac{2M}{v} \rangle$, and
- Kinetic inductance $L_k \propto \frac{m^\star}{q^2 n} = \frac{h}{q^2} \frac{1}{\langle 2Mv \rangle}$.

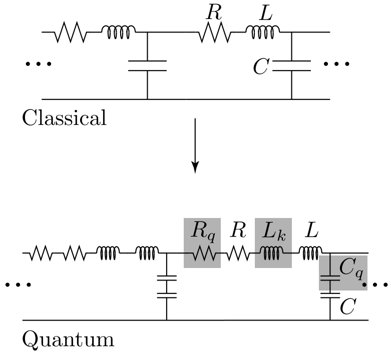


Fig. 17.1 A classical transmission line showing the resistor R, the inductor L, and the capacitor C, and the additional elements in the quantum version. In this chapter, we explore the quantum versions of these three passive circuit parameters for semiconductors.

We connect each of the these three concepts to the **electronic bandstructure** $E(k)$ and the **number of conducting modes** M of the semiconductor. It is important to realize that in a typical circuit the R, L, C is not just the three quantum parts discussed in this chapter, but includes the classical values either in series or in parallel as shown in Fig. 17.1. The point of this chapter is to emphasize that for modern electronic devices, one needs to consider the quantum counterpart to get the physics right. Though in the past the quantum components were negligible, they will not remain so in the future.

17.1 Classical R, L, C circuits

A cylindrical conductor of length L_0, diameter d, a distance h from a ground plane has a classical resistance $R = \frac{\rho}{A} \cdot L_0$, capacitance $C = \frac{2\pi\epsilon_0}{\ln(h/d)} \cdot L_0$, and inductance $L = \frac{\mu_0 \ln(h/d)}{2\pi} \cdot L_0$, relations that may be found in a textbook on electromagnetism or electronic circuits. The linear increase in these three classical quantities with length defines the per unit length quantities $\mathcal{R} = \frac{R}{L_0} = \frac{\rho}{A}$, $\mathcal{C} = \frac{C}{L_0} = \frac{2\pi\epsilon_0}{\ln(h/d)}$, and $\mathcal{L} = \frac{L}{L_0} = \frac{\mu_0 \ln(h/d)}{2\pi}$. The product $\mathcal{L}\mathcal{C} = \epsilon_0\mu_0 = \frac{1}{c^2}$ is inverse of the square of the speed of light $c = 3 \cdot 10^8$ m/s.

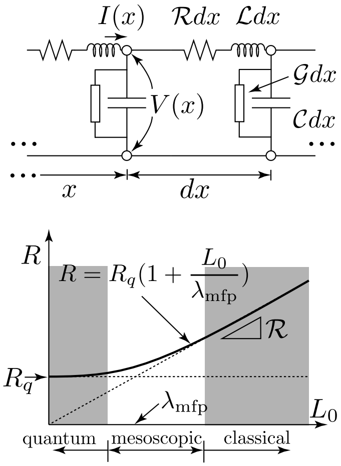


Fig. 17.2 A differential section of a transmission line showing the classical resistance, inductance, capacitance, and conductance, each of which scale linearly with the length of the section. The second figure shows that when resistance is measured down to the nanometer scales, it stops scaling linearly below roughly a mean free path of the electron.

Now imagine connecting a voltage source that applies a voltage $V(t) = V_0 e^{i\omega t}$ and injects a current $I(t) = I_0 e^{i\omega t}$ at a circular frequency $\omega = 2\pi f$ into the left side of a long conductor, which is modeled as a *RLC* network shown in Fig. 17.1. The voltage and current signals propagate to the right, and will have a spatial distribution $V(x)$ and $I(x)$ at a distance x from the voltage source. To find this variation, consider Fig. 17.2, which shows a differential element of the wire of length dx, a distance x from the source on the left. The voltage drop across the length ∂x is $\partial V(x) = -(\mathcal{R} + i\omega\mathcal{L}) \cdot \partial x \cdot I(x)$. We allow for the possibility of some current leakage from the wire to the ground through the conductance per unit length $\mathcal{G}$ in parallel with the capacitance. The change in current across dx is $\partial I(x) = -(\mathcal{G} + i\omega\mathcal{C}) \cdot \partial x \cdot V(x)$. Combining these, we obtain the celebrated **telegrapher's equation**:

$$\boxed{\frac{\partial V(x)}{\partial x} = -(\mathcal{R} + i\omega\mathcal{L}) \cdot I(x)} \text{ and } \boxed{\frac{\partial I(x)}{\partial x} = -(\mathcal{G} + i\omega\mathcal{C}) \cdot V(x)}$$

$$\implies \boxed{\frac{\partial^2 V(x)}{\partial x^2} = \gamma^2 \cdot V(x)} \text{ and } \boxed{\frac{\partial^2 I(x)}{\partial x^2} = \gamma^2 \cdot I(x)}$$

$$\boxed{\gamma = \sqrt{(\mathcal{R} + i\omega\mathcal{L}) \cdot (\mathcal{G} + i\omega\mathcal{C})}} \implies \gamma_0^2 = -\omega^2\mathcal{L}\mathcal{C} = -(\frac{\omega}{c})^2$$

$$\boxed{V(x) = V_+ e^{-\gamma x} + V_- e^{+\gamma x}} \text{ and } \boxed{I(x) = \frac{1}{Z}(V_+ e^{-\gamma x} - V_- e^{+\gamma x})}$$

$$Z = \sqrt{\frac{\mathcal{R} + i\omega\mathcal{L}}{\mathcal{G} + i\omega\mathcal{C}}}, \tag{17.1}$$

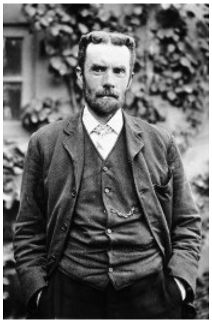


Fig. 17.3 Oliver Heaviside (1850-1925), a self-educated electrical engineer, developed the telegrapher's equation as the current-voltage version of Maxwell's equations. He introduced the mathematical concepts of the unit-step (or Heaviside) function and the impulse (or Dirac-delta) function. Heaviside was quite enamored by the power of Maxwell's equations, which in its original formulation had too many variables and symbols. The neat four-equation modern formulation with divergences and curls of **E** and **B** fields is actually due to Heaviside. He coined the word "inductance".

which is how electrical engineers disguise Maxwell's equations, as first formulated by Heaviside (Fig. 17.3). The spatial variation of the real values of voltage and current at x at any time t is obtained from the complex solution by $V(x,t) = \mathrm{Re}[V(x) \cdot e^{i\omega t}]$. The electrical signal of the source propagates as an electromagnetic wave with wavevector γ, which for the lossless transmission line $\mathcal{R} = \mathcal{G} = 0$ in vacuum is $\gamma_0 = i\frac{\omega}{c} = i\frac{2\pi}{\lambda}$, implying **the signal propagates at the speed of light in the circuit**, much faster than electrons can move, whose group velocities in crystals are limited to $v_g^{max} \leq 10^6$ m/s, $\sim 300\times$ slower than light. The ratio of the voltage and current is a characteristic impedance.

For the lossless case it is[1] $Z = \sqrt{\frac{\mathcal{L}}{\mathcal{C}}} = \sqrt{\frac{\mu_0}{\epsilon_0}} \cdot \frac{\ln(h/d)}{2\pi}$. The lower part of Fig. 17.2 motivates the purpose of this chapter: the **resistance is found not to scale linearly with the length** if the conductor length becomes smaller than the electron mean free path λ_{mfp} for scattering. As indicated in Fig. 17.1, this experimental fact, and the analogous limits of capacitance and inductance necessitates adding the quantum versions of R, L, and C to the transmission line model to get the physics right. We now discuss them, starting with quantization of conductance.

17.2 Quantized conductance

Since the classical resistance is $R = \rho \cdot \frac{L_0}{A}$, in the 1980s ultrapure semiconductor heterostructures and microscale fabrication techniques made it experimentally possible to answer a vigorously debated question: what would happen in the limit of very low impurities when $\rho \to 0$, and very small device structures when $L_0 \to 0$? If one removed all scattering, will the resistance approach zero, and conductance to ∞? There were several competing theoretical expectations, till the experimental observation shown in Fig. 17.4 cleared the air.

Fig. 17.4 shows a modulation-doped semiconductor heterostructure formed between AlGaAs and GaAs with a type-I band alignment. Donor atoms are placed in the AlGaAs layer with a spacer layer, and the GaAs is undoped. Based on the physics discussed in Section 15.6, a 2D electron gas (2DEG) forms at the heterojunction. Because the mobile electrons are physically separated from the ionized dopant scatterers, there is a significant boost in the mobility, especially at low temperatures. For example, the electron mobility measured in this experiment was $\mu = 8.5 \times 10^5$ cm^2/V·s at $T = 0.6$ K, at a 2DEG density $n_{2d} = 3.6 \times 10^{11}/\text{cm}^2$. In Chapters 21 and 23, we will quantitatively evaluate the mobilities; here, we focus on the high-mobility, or the *ballistic* limit. When the semiconductor heterostructure was processed with split gates on the surface, and the current through the 2DEG measured as a function of the voltage applied on the gate, the resistance was found to go through a series of well-defined **steps**. The conductance, as seen in the bottom panel of Fig. 17.4 was found to be quantized in units of $G = M \times (2q^2)/h$, where M is an integer. The value of the integer M changes with the gate voltage, whereas the magnitude of the electron charge q and Planck's constant h are fundamental constants. This measurement has since then been repeated in several semiconductors: in silicon, InGaAs, GaN and others, and is a *universal* feature in high mobility semiconductors. It is the quintessential example of **quantized conductance** in the ballistic limit of transport.

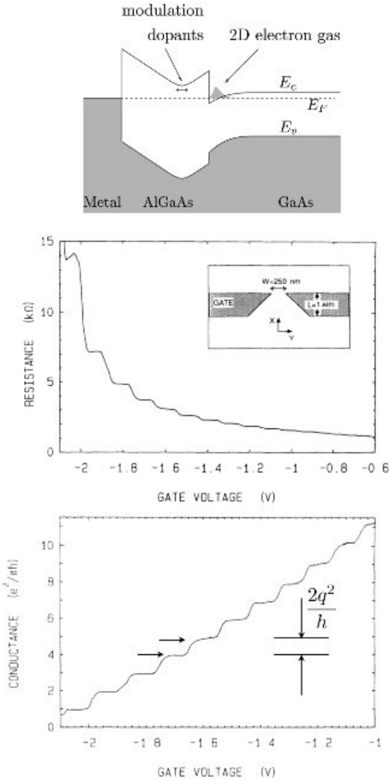


Fig. 17.4 The quantization of conductance in high mobility semiconductor heterostructure point contacts, adapted from Phys. Rev. Lett. 60, 848 (1988). Note that the conductance is quantized at $\mathbf{B} = 0$, making this form of conductance quantization different from the quantum hall effect, though there are some similar looking features. The differences will be discussed in Chapter 25.

Fig. 17.5 shows the evolution of experimentally measured conductance over nearly two centuries, from Ohm's law to ballistic conductors. In late 1960s, Sharvin had observed that the conductance became independent of the length when the length of the conductor

[1] We have treated the specific case of a cylindrical conductor here. The impedance and other characteristic quantities of the transmission line depend on the specific geometries. The failure of the lumped circuit model also occurs for high frequencies at which the electromagnetic wavelength is comparable to the size of the R, L, or C.

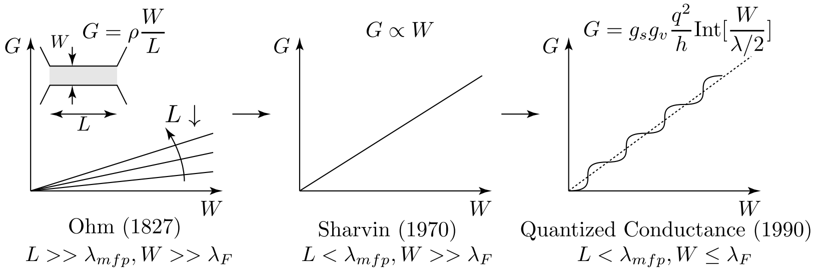

G
W
$G = \rho \frac{W}{L}$
L
$L \downarrow$
W
Ohm (1827)
$L >> \lambda_{mfp}, W >> \lambda_F$
G
$G \propto W$
W
Sharvin (1970)
$L < \lambda_{mfp}, W >> \lambda_F$
G
$G = g_s g_v \frac{q^2}{h} \text{Int}[\frac{W}{\lambda/2}]$
W
Quantized Conductance (1990)
$L < \lambda_{mfp}, W \leq \lambda_F$

Fig. 17.5 The classical "Ohm's" law 2D conductance $G = \sigma \frac{W}{L}$ was found to fail in explaining the lack of dependence on length when the distance between contacts is smaller than the mean free path of electrons (Sharvin, 1970). Then, when the electron wavelength is of the order of the width, quantized conductance is observed (1988, Fig. 17.4).

was smaller than an electron mean free path, but it remained proportional to the width, or cross-sectional area of the conductor. When the width of the conductor became of the order of the electron wavelength, only $M = \text{Int}[\frac{W}{\lambda/2}]$ modes whose wavelengths λ fit in the point contact width can transmit, the others must reflect. For a Fermi wavevector $k_F = 2\pi/\lambda_F$, the number of modes are $M = \text{Int}[\frac{k_F W}{\pi}]$ in 2D and $M = \text{Int}[\frac{k_F^2 A}{\pi^2}]$ in 3D. This picture of conductance is similar to sound or light wave propagation in waveguides, as shown in Fig. 17.6.

That the conductance is quantized in units of $2q^2/h$ should not be that big of a surprise to you: in Chapter 5, Equation 5.40 we learned that this must be the case for a 1D conductor with free–electrons, in the absence of scattering[2]. The group velocity of a mode k of the energy band is $\mathbf{v}_g = \hbar^{-1} \nabla_k E(k)$, and therefore the current

$$I = \frac{q g_s g_v}{L} \sum_k v_g(k) f(k) = \frac{q g_s g_v}{\hbar L} \int_{-\infty}^{+\infty} \frac{dk}{\frac{2\pi}{L}} \frac{\partial E(k)}{\partial k} f(k)$$

$$I = \frac{q g_s g_v}{2\pi\hbar} (k_b T) \ln \left(\frac{1 + e^{\frac{E_{Fs} - E_F}{k_b T}}}{1 + e^{\frac{E_{Fd} - E_F}{k_b T}}} \right) \implies @T = 0 \rightarrow \frac{I}{V} = g_s g_v \frac{q^2}{h} \quad (17.2)$$

is the product of the velocity of the modes and the occupation function, summed over all k-states. In the ideal 1D situation, the number of cross-sectional modes is by definition $M = 1$, the spin-degeneracy produces $g_s = 2$ states of equal energy, and the valley degeneracy g_v due to the crystal structure can produce further copies of equal-energy states at different k points in the Brillouin zone. Fig. 17.6 shows the electron (and hole) bandstructure of the conductor. The occupation function is controlled by the electrochemical potentials $\mu_1 = E_{Fs}$ and $\mu_2 = E_{Fd}$ of the source and the drain contacts, and the difference is $\mu_1 - \mu_2 = E_{Fs} - E_{Fd} = qV$, where V is the applied voltage. For 1D, the group velocity cancels with the density of states, leading to conductance quantization that is independent of the exact nature of the bandstructure $E(k)$. As the split gate voltage is swept, lateral depletion changes the electronic width of the constriction, changing the number of 1D modes that fit the width. Each mode that fits contributes a conductance quantum, explaining the steps observed in Fig. 17.4. We can

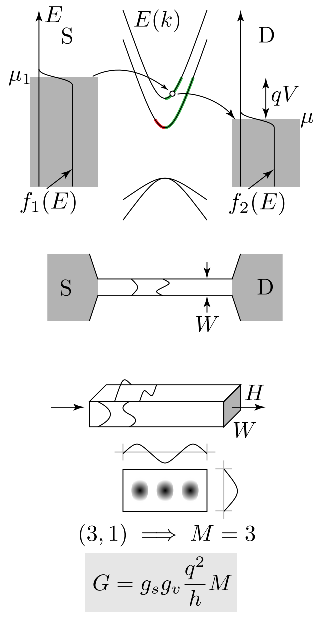


Fig. 17.6 Current flow driven by the Fermi level difference by filling states from the source contact and emptying states into the drain contact. The maximum conductance is limited by the number $M = \text{Int}[\frac{W}{\lambda/2}]$ or $\text{Int}[\frac{WH}{(\lambda_x/2)\cdot(\lambda_y/2)}]$ of electron half-wavelengths that fit in the cross-section of the conductor, each providing a conductance of q^2/h.

[2] The quantization of conductance, though a revelation at the time – was observed 8 years *after* the observation of the integer quantum Hall effect, in which very precise conductance quantization, also in units of Nq^2/h is measured in 2DEGs. However, the Hall conductance is the *transverse* conductance σ_{xy} in the presence of a magnetic field, whereas what we are discussing here is the standard two-terminal conductance σ_{xx} or resistance with no magnetic field. The quantum Hall effect is discussed in Chapter 25.

then generalize the ballistic conductance for different dimensions:

$$\boxed{G_q = g_s g_v \frac{q^2}{h} M}, \; M = \underbrace{1}_{1\text{D}}, = \underbrace{\text{Int}[\frac{W}{\lambda_F/2}]}_{2\text{D}}, = \underbrace{\text{Int}[\frac{W}{\lambda_{Fx}/2} \cdot \frac{H}{\lambda_{Fy}/2}]}_{3\text{D}}, \tag{17.3}$$

where λ_F is the Fermi wavelength of electrons in the highest filled states. The cross-section into which the half-wavelengths must fit is of one lower dimensionality than the conductor itself. This viewpoint is consistent with the ballistic current density we had obtained in Chapter 5 for a conductor with parabolic bands in d-dimensions, where the lower dimensional effective DOS N_c^{d-1} appears explicitly:

$$J_d = \frac{q^2}{h} \cdot N_c^{d-1} \cdot \frac{k_b T}{q} \cdot [F_{\frac{d-1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) - F_{\frac{d-1}{2}}(\frac{E_{Fs} - E_c - qV}{k_b T})]. \tag{17.4}$$

We discuss the experimental details of the step-like change of conductance to bring out the important interplay of scattering, electron mean free paths, quasi-Fermi levels, and non-equilibrium physics that will be central to electron transport phenomena discussed in later chapters. Landauer (Fig. 17.7) had pioneered viewing electron transport not as the Drude picture of electrons as classical particles scattering from point defects, but as modes of electron waves that undergo reflection or transmission across the potential barriers caused by scattering sites. For a quantitative feel of length scales, the Fermi wavelength λ_F, and mean free path λ_{mfp} for a 2DEG of carrier density n_{2d} and mobility μ are:

Fig. 17.7 Rolf Landauer in 1950s at IBM realized the role of the wave nature of electrons on the electrical and thermal conductance in quantized structures, and contributed to quantifying the physics of information.

$$\frac{2\pi}{\lambda_F} = k_F = \sqrt{\frac{4\pi n_{2d}}{g_s g_v}} \implies \boxed{\lambda_F = \sqrt{\frac{\pi g_s g_v}{n_{2d}}} = \frac{25 \text{ nm}}{\sqrt{(n_{2d}/10^{12}/\text{cm}^2)}}}$$

$$\lambda_{\text{mfp}} \sim v_F \tau, \text{ where } \mu = \frac{q\tau}{m^\star}, \lambda_{\text{mfp}} \sim \frac{\hbar k_F}{m^\star} \cdot \frac{\mu m^\star}{q}$$

$$\implies \boxed{\lambda_{\text{mfp}} \sim \frac{\hbar \mu}{q} \cdot \sqrt{\frac{4\pi n_{2d}}{g_s g_v}}}$$

$$\implies \lambda_{\text{mfp}} \sim [(\frac{\mu}{10^3 \frac{\text{cm}^2}{\text{V}\cdot\text{s}}}) \cdot \sqrt{\frac{n_{2d}}{10^{12} \text{ cm}^2}} \cdot 16.5] \text{ nm} \tag{17.5}$$

where the numerical formulae are for $g_s = 2$ and $g_v = 1$, true for the GaAs heterostructure. For a starting 2DEG density $n_{2d} = 3.6 \times 10^{11}/\text{cm}^2$ used in Fig. 17.4, the Fermi wavelength is $\lambda_F \approx 40$ nm – much larger than atomic dimensions, which makes it possible for the gate voltage to control a small number of electron modes to pass through the point contact, leading to the conductance quantization. For $\mu = 8.5 \times 10^5$ cm^2/V·s, the mean-free path is $\lambda_{mfp} \sim 8.4$ μm. Fig. 17.8 indicates the range of Fermi wavelengths and mean-free paths encountered in typical semiconductor structures. Note that **the Fermi**

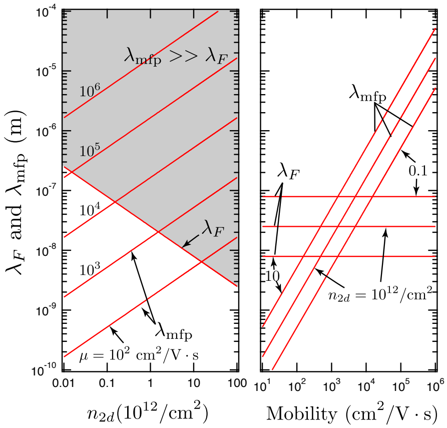


Fig. 17.8 Fermi wavelengths and mean free paths as a function of the 2D electron gas density (left) and electron mobility (right). For example, the mean free path of electrons at the Fermi level at a density of $n_{2d} = 10^{12}/\text{cm}^2$ for a mobility of 10,000 cm^2/V·s is $\lambda_{mfp} \sim$ 165 nm, about 8× larger than the Fermi wavelength of $\lambda_F \sim 25$ nm, both of which are obtained from Equation 17.5. The shaded region indicates that higher 2DEG densities enable regimes in which the mean free path is larger than the Fermi wavelengths. Note that while the mean free path λ_{mfp} depends on the mobility and bandstructure, the Fermi wavelength λ_F does not: it is purely a function of the electron density.

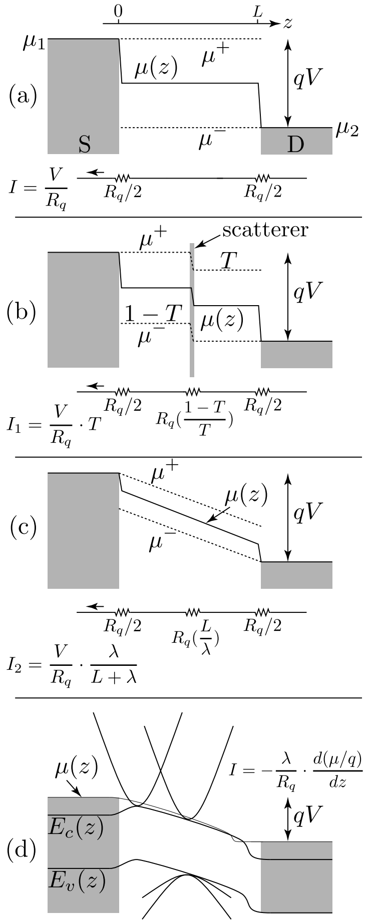


Fig. 17.9 The evolution of energy diagrams showing the evolution of the spatial variation of quasi-Fermi levels from the ballistic (a) to a single scatterer (b), to a quasi-diffusive (c), to a completely diffusive (d) regime of transport. A single mode-based approach accounts for the conductance of all cases. The ballistic picture of wave-fitting leads to effective voltage drops at the contacts, and scatterers introduce additional voltage drops, as indicated by the resistor network circuit diagrams. For very long conductors, the difference of the quasi-Fermi levels is too small, and a single $\mu(z)$ as shown in (d) is used to capture the transport and electrostatics.

wavelength is independent of electron transport properties and the details of the semiconductor bandstructure, it depends solely on the density of electrons.

Fig. 17.9 quantitatively indicates how Landauer's picture of a ballistic conductor evolves to a "drift-diffusion", or classical conductor in the presence of increasing scattering in the channel. Consider the conductor of length L with no scatterers. A voltage difference of V between the contacts creates two populations of carriers in the channel: those in equilibrium with the source, which share its quasi Fermi-level $\mu^+ = \mu_1$ and constitute carriers moving to the right, from the source to the drain. The second population of electrons move to the left, and have a quasi-Fermi level $\mu^- = \mu_2$, the Fermi level of the drain. Since in the ballistic case (a) there is no scattering, the quasi-Fermi levels stay constant from the source to the drain since no right-moving carrier scatters to become a left-moving carrier, and no left-moving carrier scatters to the right. The electrochemical potential that characterizes the total carrier concentration at z is the average $\mu(z) = \frac{1}{2}(\mu^+ + \mu^-)$. The current of Equation 17.2 satisfies the current continuity and boundary conditions:

$$I = G_q(\frac{\mu^+ - \mu^-}{q}), \text{ and } \frac{dI}{dz} = 0, \text{ and}$$
$$\mu^+(z = 0^+) = \mu_1 \text{ and } \mu^-(z = L^-) = \mu_2, \tag{17.6}$$

where the ballistic conductance G_q is given by Equation 17.3. In steady state, current continuity implies $I =$constant in z. For the ballistic case of Fig. 17.9 (a), $\mu(z)$, the average of μ^+ and μ^-, drops abruptly at the source and drain contacts, and stays constant in the channel, implying there is no voltage drop in the channel. Writing $R_q = 1/G_q$, the voltage drops by $IR_q/2$ at each contact, **as if there were two resistors $R_q/2$ at each contact**, indicated by the circuit diagram.

Fig. 17.9 (b) shows how the situation changes upon the introduction of a single scatterer. A fraction T of the right-going carriers manage to transmit across the scatterer, whereas $1 - T$ are reflected back towards the source. This is captured in the drop in μ^+ moving to the right across the scatterer, and the corresponding increase in μ^- crossing the scatterer to the left. The physical meaning is clear: the electrons that are reflected back to the left from the scatterer increase the left-going quasi-Fermi level. The drop in the total electrochemical potential across the scatterer indicates that it introduces an additional resistance $R_q(\frac{1-T}{T})$. Fig. 17.9 (c) shows a "zoomed-out" picture for the case when a large number of such scatterers are introduced throughout the channel. Writing the mean-free path as $\lambda_{mfp} = \lambda$, the net resistance in the channel becomes $R_q\frac{L}{\lambda}$, related to the single scatterer of part (b) via $T = \frac{\lambda}{L+\lambda}$, if coherent scattering and interference effects are neglected[3].

[3] Coherent scattering and interference effects can lead to resonant tunneling and other exotic phenomena, as discussed in Chapter 24.

Thus for both the diffusive and ballistic regimes,

$$\mu^+ - \mu^- = \frac{\lambda}{L+\lambda}\cdot(\mu_1 - \mu_2)$$

$$\frac{\mu^+ - \mu^-}{q} = IR_q, \text{ and } I = G_q\left(\frac{\mu^+ - \mu^-}{q}\right) = \frac{G_q\lambda}{L+\lambda}\left(\frac{\mu_1 - \mu_2}{q}\right). \tag{17.7}$$

In the diffusive limit, when there is a gradient in the electrochemical potential as seen in Figures 17.9 (c) and (d), the current is equivalently found from the equation

$$I = -\frac{G_q\lambda}{q}\frac{d\mu}{dz} = -\frac{\sigma A}{q}\frac{d\mu}{dz} \approx \frac{G_q\lambda}{L+\lambda}\left(\frac{\mu_1 - \mu_2}{q}\right) = \frac{V}{R_q(1+\frac{L}{\lambda})}, \tag{17.8}$$

which is useful especially for very long channels when $\mu^+ - \mu^- = \frac{\lambda}{L+\lambda}(qV) \approx 0$, and the $I = G_q(\mu^+ - \mu^-)/q$ form, though still correct, is not practical to use. Fig. 17.9 (d) shows a long-channel FET that we will encounter in Chapter 19, for which the separation between μ^+ and μ^- is negligible, yet $\mu(z) = (\mu^+ + \mu^-)/2$ shown has a spatial variation distinct from the band-edge $E_c(z)$. The current is found from the gradient of $\mu(z)$, not $E_c(z)$.

To summarize[4], accounting for the quantized conductance makes the resistance of a conductor $R(L) = R_q(1 + L/\lambda)$. When $L \to 0$, the resistance approaches R_q as was indicated in Fig. 17.2 for small conductors. Whether one should use the diffusive limit of Equation 17.8 or ballistic (or quasi-ballistic) limit of Equation 17.7 depends on the problem. The concept of quasi-Fermi levels will prove to be of great importance in following chapters. Here, we encountered the quasi-Fermi levels of right-going and left-going carriers. In Chapter 18 we will use quasi-Fermi levels to understand the operation of ballistic Schottky diodes. We will then introduce two quasi-Fermi levels, one for electrons in the conduction band and the other for holes in the valence band to understand the operation of pn junctions, and bipolar transistors in the diffusive transport regime. In Chapter 19 we will use quasi-Fermi levels for both ballistic and diffusive regimes to understand the operation of field-effect transistors. In Chapter 21 we will describe the behavior of the quasi-Fermi levels with the Boltzmann transport equation, and use them in Chapters 27, 28, and 29 to understand light-matter interaction, and photonic device operation.

How does the conductance scale down to the last atoms? If there was a single atom, or its semiconductor counterpart, a heterostructure quantum dot, connected to a source reservoir at Fermi level E_{Fs} and a drain reservoir at E_{Fd}, the maximum conductance for an electron to transport through a single *discrete* level of a quantum dot is also $G = 2q^2/h$. The reason for this is highlighted in Fig. 17.10. The coupling of the discrete energy eigenvalue to the electrodes causes an effective broadening of the energy level, which is a way of saying that the energy is not a state of definite energy since an electron put in it can leak out to the drain. A single energy level can lead to on-site Coulomb

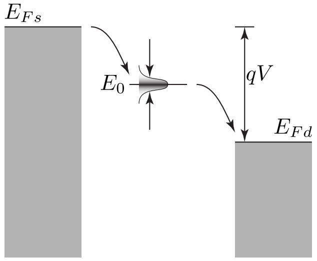


Fig. 17.10 Conductance between two electrodes through a single atomic level, or a quantum dot. The maximum conductance is $2q^2/h$. Due to the perturbation of the eigenvalue of the quantum dot by wavefunction overlap with those of the two electrodes, the sharp eigenvalue of the isolated dot is broadened, making it not a state of definite energy, and allowing transport through it.

---

[4] Equations 17.2 and 17.3, together with 17.7 and 17.8 convey the Landauer picture of conductance when the role of inelastic scattering (e.g. by phonons) is neglected. The role of elastic scatterers was first identified by Landauer, and those of the contacts by Yoseph Imry. Our earlier discussion in Chapter 5 of conductance quanta and current saturation in the ballistic regime of transport for **free–electrons** carries over to transport in crystals directly, with the free–electron mass replaced by the band-edge effective masses, and the zero energy by the band edges. Exercise 17.2 gives you an opportunity to practice and become comfortable in this way of thinking. In 1D, the effective mass does not appear explicitly, but it does in 2D and 3D transport. The 2D and 3D ballistic transport cases may be derived by considering them as parallel arrays of 1D conductors, and using the simple conductance quantum $2q^2/h$ for each mode with the correct Fermi level differences.

repulsion and many-particle effects, similar to the donor and acceptor dopants discussed in Chapter 15. Though a 2D conductor is the most popular, and 1D versions are increasingly used for efficient transport of carriers in transistors, the quantum dot version is also increasing in importance in several applications ranging from memories, to LEDs and lasers.

17.3 Quantum capacitance

The capacitance C of a device is a measure of its capacity to store electrical energy U in the form of charge Q. The classical capacitance of a conductor depends solely on its **geometry**, because the electrical energy is stored in the small displacements of the electron clouds in the insulating, or dielectric regions in the space outside the conductor(s). Since the DC electric field inside a perfect conductor is zero, no energy is stored inside the conductor for the classical capacitor. This classical picture is incomplete in nanoscale capacitors: a part of the electrical energy is also stored inside the conductor. This property of a conductor is called its **quantum capacitance**. The electrical energy divides between the quantum and classical geometric capacitances.

To add a charge dQ, a voltage dV must drop across the electrodes connected to the device. The ratio $C = dQ/dV$ is defined as its capacitance. The total charge Q is related to the total voltage drop V by $\int dQ = C \int dV$. To put a total charge $Q = CV$ on the device, an energy

$$\begin{aligned} U = \int_{V=0}^{V} Q' dV' = \int_{V=0}^{V} CV' dV' = \frac{1}{2} CV^2, \\ \text{or } U = \int_{V=0}^{V} Q' dV = \int_{Q=0}^{Q} \frac{Q'}{C} dQ' = \frac{Q^2}{2C} \end{aligned} \tag{17.9}$$

must be provided to the capacitor. Experimentally a battery provides the external voltage V, and the charges Q respond to this voltage. Since $U = \frac{1}{2}CV^2$, to store a larger amount of energy for the same voltage drop, a higher capacitance[5] is desired. As standard textbook examples, the classical capacitance of a parallel plate capacitor of thickness d and area A is $C_{\|} = \epsilon_0 \kappa \frac{A}{d}$, $C_{sph} = 4\pi\epsilon_0 \kappa R$ for a sphere of radius R, and is $C_{cyl} = \frac{2\pi\epsilon_0\kappa}{\ln(h/d)} L$ for a cylindrical conductor of length L, and diameter d a distance h from a ground plane.

To find the quantum capacitance of a conductor, we first note that electrons occupy allowed energy eigenvalues $E(k)$ of the conductor up to the Fermi level, or an electrochemical potential μ, with a Fermi–Dirac distribution f_k in energy. The change of the total charge $Q = q \sum_{\mathbf{k}} f_k$

---

[5] High-κ dielectric materials enable higher capacitance, as may be directly seen for the parallel plate capacitor $C = \epsilon_0 \kappa \frac{A}{d}$ of thickness d and area A. Viewing a capacitor as a device that stores *energy*, high-κ materials enable compact batteries. Viewing it as a device that stores *charge* and helps manipulate charge with better energy efficiency, high-κ dielectrics are desired for both semiconductor memories, and in field-effect transistors, the logic element.

with the electrochemical potential is the quantum capacitance:

$$f_k = \frac{1}{1+e^{\frac{E(k)-\mu}{k_bT}}}, Q = q\sum_{\mathbf{k}} f_k \implies \frac{\partial Q}{\partial(\mu/q)} = q^2 \sum_{\mathbf{k}} \underbrace{\frac{\partial f_k}{\partial \mu}}_{\frac{1}{4k_bT\cosh^2[\frac{E(k)-\mu}{2k_bT}]}}$$

$$\implies \boxed{C_q = \frac{q^2}{4k_bT}\int \frac{d^dk}{(\frac{2\pi}{L})^d}\cdot\frac{1}{\cosh^2[\frac{E(k)-\mu}{2k_bT}]} \underbrace{\approx}_{T\to 0} q^2\cdot g(\mu)\cdot L^d.} \tag{17.10}$$

This general formulation provides the quantum capacitance C_q of any conductor whose bandstructure $E(k)$ is known. The resulting C_q per unit "volume" is found to be $q^2 \times g(\mu)$, i.e., the electron charge squared times the DOS at the Fermi level[6]. For example, for metallic carbon nanotubes, $E = \hbar v_F|k|$, and

$$C_q = \frac{q^2 g_s g_v}{4k_bT}\cdot\int_{-\infty}^{+\infty}\frac{dk}{(\frac{2\pi}{L})}\cdot\frac{1}{\cosh^2[\frac{\hbar v_F|k|-\mu}{2k_bT}]} = L\cdot\underbrace{\frac{2g_sg_vq^2}{hv_F}}_{q^2\cdot\text{DOS}}\cdot\underbrace{\frac{1}{1+e^{-\frac{\mu}{k_bT}}}}_{\mu>>k_bT\implies\approx 1}. \tag{17.11}$$

Similarly, for a 2DEG with a parabolic bandstructure $E(k_x,k_y) = E_c + \frac{\hbar^2}{2m_c^\star}(k_x^2+k_y^2)$, the quantum capacitance evaluates to

$$\begin{aligned} C_q &= \frac{q^2g_sg_v}{4k_bT}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{dk_x}{(\frac{2\pi}{L_x})}\frac{dk_y}{(\frac{2\pi}{L_y})}\cdot\frac{1}{\cosh^2[\frac{E_c+\frac{\hbar^2(k_x^2+k_y^2)}{2m_c^\star}-\mu}{2k_bT}]} \\ &= \frac{L_xL_yq^2g_sg_v}{4k_bT}\int_{\theta=0}^{2\pi}\int_{k=0}^{+\infty}\frac{kdkd\theta}{(2\pi)^2}\cdot\frac{1}{\cosh^2[\frac{\frac{\hbar^2k^2}{2m_c^\star}-(\mu-E_c)}{2k_bT}]} \\ &= L_xL_y\cdot\underbrace{\frac{q^2g_sg_vm_c^\star}{2\pi\hbar^2}}_{q^2\cdot\text{DOS}}\cdot\frac{1}{1+e^{-(\frac{\mu-E_c}{k_bT})}}, \end{aligned} \tag{17.12}$$

which for $\mu - E_c >> k_bT$ again yields the general form $C_q/A = q^2\cdot$ DOS. As long as the Fermi level is a few k_bT inside a conducting band (either conduction or valence band)[7], the semiconductor is degenerate, the quantum capacitance is faithfully reproduced by $C_q/A = q^2\cdot\text{DOS}$.

How is the capacitance C related to the atoms and materials and their geometry that make the device? Microscopically, where exactly is this energy stored? For the classical, or geometric capacitance **the electrical energy $U = \frac{1}{2}CV^2$ is stored in the displaced negatively charged electron clouds with respect to the positively charged nuclei of the insulating dielectric, or material medium filling the space between the conductors**. The displacement is due to the force exerted on the charges by the external electric field, which perturbs them from their equilibrium positions. But there is a bit more energy stored in the electrons inside the (semi)conductor too, that is typically neglected. This portion of the energy is a consequence of the Pauli exclusion principle, and the capacitance it defines is therefore called the **quantum capacitance**. The general form of the quantum capacitance as derived above is $C_q = q^2 \cdot \text{DOS}$, that is, the square of the electron charge times the density of states at the Fermi level[8]. Let us first see how this result is obtained quantitatively from a 2D parallel plate capacitor model by drawing the energy-band diagram.

---

[6] See Exercise 17.3 for an equivalent formulation of quantum capacitance in terms of conducting modes, $C_q = \frac{q^2}{h}\langle\frac{2M}{v}\rangle$.

[7] Note that when $\frac{\mu-E_c}{k_bT} << -1$, meaning when the Fermi level is deep inside the bandgap, $C_q \to 0$. Thus the quantum capacitance changes from exponentially small to C_q as $\frac{\mu-E_c}{k_bT}$ sweeps though zero.

Fig. 17.11 shows a parallel plate capacitor in the middle, connected to a voltage source that applies a voltage V. The capacitor is formed between a metal electrode plate, a thin insulator layer of barrier thickness t_b and dielectric constant ϵ_b, and a second electrode. If the second electrode is also a metal plate, the corresponding charge $\rho(z)$, electric field $F(z)$, and energy band diagrams are shown to the left. If the second electrode is a *semiconductor*, the corresponding quantities are shown to the right[9].

Let us first discuss the simpler case of the metal-insulator-metal (M-I-M) capacitor. For a positive voltage V on the voltage source, electrons are extracted from the top plate and injected into the bottom plate. The top electrode is now positively charged, and the bottom electrode holds a net negative charge, equal in magnitude to the top electrode: the battery[10] is always charge neutral, and so is the M-I-M capacitor. Since a metal does not allow DC electric fields in its interior, all the charges reside on the surface of the metal adjacent to the insulator. On the top electrode, the positive charges are the metal ions next to the insulator that have lost their electrons, and on the bottom electrode, the negative charges are the extra electrons that entered it, located next to the insulator. The thickness of these charged layers are roughly one lattice constant, which allows us to model the charge as a 2D sheet charge, of density $q\sigma_m$. By Gauss's law, there must now be a *constant* electric field $F_b = \frac{q\sigma_m}{\epsilon_b}$ pointing from the top to the bottom electrode. The area under the electric field $\int dz \cdot F(z) = V$ is exactly equal to the voltage of the battery. For a cross section area A we have

$$V = \frac{q\sigma_m}{\epsilon_b} t_b \implies Q = A \cdot q\sigma_m = (\epsilon_b \frac{A}{t_b}) \cdot V = C \cdot V, \tag{17.13}$$

and we recover the standard expression for a parallel plate capacitor $C = AC_b$, where $C_b = \frac{\epsilon_b}{t_b}$ is the capacitance per unit area. The energy band diagram of the M-I-M capacitor on the left shows that the field causes a bending of the bands of the insulator, which is a material of a large bandgap E_b. The Fermi level difference between the top and bottom metal electrodes is qV. The energy $U = \frac{1}{2}CV^2$ given to the capacitor from the battery is stored as the potential drop across the insulator[11].

If the bottom electrode is replaced by a semiconductor, we have a metal-insulator-semiconductor (M-I-S) capacitor. To illustrate the idea of the quantum capacitance, we assume that the semiconductor

---

[8] This result holds as long as many-particle effects do not dominate the energies. If the capacitor involves quantum dots, electron-electron interactions can significantly alter the capacitance, which will be discussed shortly.

[9] The difference in the work functions of the two electrodes defines a *threshold voltage*, of great importance in diodes and transistors.

[10] We will use the terms voltage source and battery interchangeably.

[11] Microscopically, the centroid of the atomic electron clouds in the insulating barrier coincide with the atomic nuclei before the voltage was applied. The electric field resulting from the voltage stretches the electron cloud preferentially to one direction, forming an electric dipole. The energy U is stored in these electric dipoles, which may be imagined as stretched springs (see Fig. 26.5 for an illustration). A dielectric material in which the electrons are bound strongly to the nuclei has a small dielectric constant, and one in which electrons are bound weakly has a high ϵ_b.

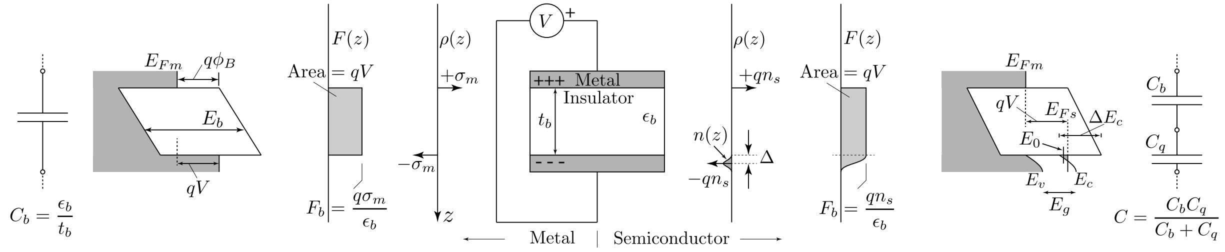


Fig. 17.11 Illustrating the origin of quantum capacitance C_q due to the spread of electrons in the semiconductor.

is a thin membrane[12], say of 10 nm thickness, doped n-type heavily putting the Fermi level close to the conduction band edge, and has ohmic contacts to the conduction band. Unlike a metal, the *excess* electrons injected into the semiconductor cannot be within a lattice constant of the I-S interface, because of the far smaller DOS at the conduction band edge. The excess electrons must distribute in volume over a larger spatial distance $n(z)$ as indicated in Fig. 17.11, leading to a finite field inside the semiconductor layer falling rapidly away from the I-S interface. The net *sheet* charge density in the semiconductor $-qn_s = -q\int dz \cdot n(z)$ must be balanced by $+qn_s$ in the metal. Unlike the M-I-M capacitor, the electric field penetrates partially into the semiconductor, and yet the area under the $F(z)$ curve must still be V. Therefore, for the same voltage as the M-I-M capacitor, the electric field in the barrier $F_b = \frac{qn_s}{\epsilon_b}$ must then be *smaller* in the M-I-S capacitor. Another way to arrive at the same conclusion is to see that the centroid of the electron distribution in the semiconductor cannot be at the I-S junction, but a few lattice constants Δ away from it, as indicated in the figure. This implies a smaller parallel plate capacitance approximated as $C = \epsilon_b \cdot \frac{A}{t_b + \Delta}$ if the insulator and the semiconductor share the same dielectric constant. This difference is the root of the quantum capacitance, which can be viewed as an additional capacitor C_q in series with the standard barrier capacitance C_b. Let us now quantitatively calculate the quantum capacitance.

The method of calculation is to find the dependence of the sheet density n_s on the voltage V. To do so, we can directly write, from the M-I-S energy band diagram, the energy conservation equation from the metal to the semiconductor:

$$E_{Fm} + q\phi_b + q\underbrace{F_b t_b}_{V_b} - \Delta E_c + E_0 + (\underbrace{E_{Fs} - E_0}_{f(n_s)}) = E_{Fs}. \tag{17.14}$$

We identify the difference $E_{Fs} - E_{Fm} = qV$, and club the material constants that do not depend on the voltage to define a characteristic voltage V_T via $V_T = \phi_b - \frac{\Delta E_c}{q} + \frac{E_0}{q}$. Since $F_b = \frac{qn_s}{\epsilon_b}$ and $C_b = \frac{\epsilon_b}{t_b}$, we get

$$\frac{q^2 n_s}{C_b} + (E_{Fs} - E_0) = q(V - V_T). \tag{17.15}$$

[12] The semiconductor can be a 3D, 2D, or a 1D crystal. The arguments presented for the 2D case carry over with appropriate dimensional adjustments.

The excess electrons form a 2D electron gas in the semiconductor. We assume here that one subband of energy minimum E_0 is occupied[13]. The location of the Fermi level in the semiconductor is related to the electron density via $n_s = n_{2d} = N_c^{2d} F_0(\frac{E_{Fs}-E_0}{k_bT})$, where $F_0(\eta) = \ln(1+e^{\eta})$ is the zeroth order Fermi–Dirac integral. Thus, we have

$$n_s = N_c^{2d} \ln(1 + e^{\frac{E_{Fs}-E_0}{k_bT}}) \implies e^{\frac{E_{Fs}-E_0}{k_bT}} = e^{\frac{n_s}{N_c^{2d}}} - 1. \tag{17.16}$$

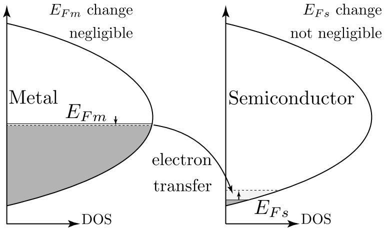


Fig. 17.12 When electrons are transferred from a metal to a semiconductor, the change in the metal Fermi level is negligible compared to that in the semiconductor. This is because the metal has a far larger DOS at the Fermi level than the semiconductor. The extra voltage drop in the semiconductor to accommodate the charge is the origin of the quantum capacitance.

This suggests dividing Equation 17.15 by k_bT throughout to make the terms dimensionless. Defining $\frac{k_bT}{q} = V_{th}$, $n_b = \frac{C_b V_{th}}{q}$, we exponentiate both dimensionless sides of the equation to obtain

$$\boxed{e^{\frac{n_s}{n_b}} \cdot (e^{\frac{n_s}{N_c^{2d}}} - 1) = e^{\frac{V-V_T}{V_{th}}}.} \tag{17.17}$$

The dependence of n_s on V is seen explicitly in this transcendental equation. We can find two limits of the sheet density: if $n_s << N_c^{2d}$, $e^{n_s/N_c^{2d}} - 1 \approx n_s/N_c^{2d}$, and to leading order, $n_s \approx N_c^{2d} e^{\frac{V-V_T}{V_{th}}}$, implying the electron density in the semiconductor is exponentially dependent on the voltage. This situation is obtained when $V < V_T$, and the quasi-Fermi level of the semiconductor $E_{Fs} < E_0$ is inside the bandgap of the semiconductor, providing very few electrons. On the other hand, if $V > V_T$, the Fermi level rises to $E_{Fs} > E_0$, the sheet density $n_s \geq N_c^{2d}$ is high, and neglecting the -1 in the LHS of Equation 17.17 we obtain

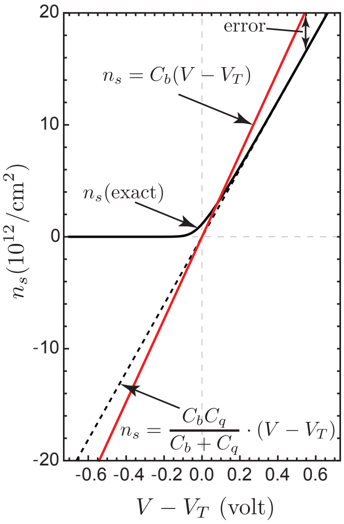


Fig. 17.13 Illustration of the errors incurred by neglecting the quantum capacitance. The values used are $\epsilon_b = 20\epsilon_0$, $t_b = 3$ nm, $m^\star = 0.2m_e$, $g_s = g_v = 2$, and $T = 300$ K. This represents for example a metal-HfO$_2$-silicon M-I-S capacitor. Note the error at high $V - V_T$ if the C_q is neglected. This regime of operation represents the on-state of a transistor if the semiconductor is used as the conducting channel.

$$(\frac{1}{n_b} + \frac{1}{N_c^{2d}})n_s \approx \frac{V-V_T}{V_{th}} \implies n_s \approx \frac{1}{(\frac{1}{n_b} + \frac{1}{N_c^{2d}})V_{th}}(V - V_T)$$

$$n_s \approx \frac{1}{(\frac{1}{C_b} + \frac{V_{th}}{N_c^{2d}})}(V - V_T) \approx \frac{1}{(\frac{1}{C_b} + \frac{1}{C_q})}(V - V_T)$$

$$\implies \boxed{n_s \approx \frac{C_b C_q}{C_b + C_q}(V - V_T).}$$

$$\implies \boxed{C_q = \frac{N_c^{2d}}{V_{th}} = q^2 \cdot \frac{g_s g_v m^\star}{2\pi\hbar^2} = q^2 \cdot \text{ DOS }.} \tag{17.18}$$

The **quantum capacitance** C_q identified here thus represents the part of the energy, or voltage drop that occurs in the semiconductor, whereas C_b represents the standard geometric capacitance. Because the DOS of the semiconductor at its Fermi level is much smaller at that of the metal, a finite voltage drop is necessary to sustain charge, as indicated in Fig. 17.12. From the expression for quantum capacitance in Equation 17.12, the higher the effective mass, the higher the DOS and hence the higher the C_q. Since C_q appears in series with C_b, **the smaller of the two dominates**. We can equate the quantum capac-

[13] This approximation may be relaxed to account for multiple subbands.

itance to an effective increase in the insulator thickness by writing

$$C_q \approx \frac{\epsilon_b}{\Delta} = q^2 \cdot \text{ DOS} \implies \Delta \approx \frac{\epsilon_b}{q^2 \cdot \text{DOS}} = \frac{2\pi\hbar^2\epsilon_b}{q^2 g_s g_v m^\star} = \frac{2}{g_s g_v} \cdot \frac{\epsilon_b}{\epsilon_s} \cdot \frac{a_B^\star}{4}, \tag{17.19}$$

where the effective barrier thickness $t_b + \Delta$ is seen to increase in proportion to the effective Bohr radius $a_B^\star = (4\pi\epsilon_s\hbar^2)/q^2 m^\star$ of the semiconductor. A smaller effective mass semiconductor therefore has a smaller C_q, and electrons in it spread over a larger distance. As semiconductor transistors are miniaturized, the spread in the electron distribution in space must be tightly controlled. Using higher effective mass semiconductors helps with this control in scaling.

Fig. 17.13 illustrates the importance of considering the quantum capacitance in semiconductor devices. If by neglecting C_q, we assume that the sheet density in the semiconductor is simply $n_s = C_b(V - V_T)/q$, a large error is incurred for $V - V_T >> k_b T$, or in the high conductivity "on" state of the semiconductor. Inclusion of the quantum capacitance via the $n_s = \frac{C_b C_q}{C_b + C_q}(V - V_T)/q$ approximation matches the exact dependence of n_s on V captured by Equation 17.17 in the "on"-state. However, neither of the approximations can capture the low conductivity "off" state for $V - V_T \lessapprox 0$, because when the Fermi level E_{Fs} enters the bandgap, the semiconductor behaves as an insulator rather than a metal[14]. If $t_b \approx \Delta$, neglecting the quantum capacitance leads to large errors, whereas for $t_b >> \Delta$ the error is negligible.

Since the DOS can change in different dimensions, the relative interplay of C_q and C_b can be significantly altered. For example, since in 1D, the DOS $g_{1d} \sim 1/\sqrt{E - E_c}$ has a van-Hove singularity at the band edge and drops rapidly, it can dominate over C_b when it becomes $C_q << C_b$. The case of 0D quantum dots is interesting because many-particle effects manifest themselves in the quantum capacitance. Let us briefly discuss this case because it highlights a subtle issue about quantum capacitance.

What is the capacitance of a single atom? Imagining the atom as a quantum dot, let E_N represent the N^{th} energy eigenvalue that has accounted for the electron-electron repulsion energy incurred by successive filling of the eigenvalues. The energy required to put an additional electron on a system that has N electrons bound in all 3 dimensions is

$$U(N+1) - U(N) = E_{N+1} + \frac{q^2}{4\pi\epsilon_0 R_{e-e}}, \tag{17.20}$$

where the Coulomb repulsion term is the energy for adding the $(N+1)^{\text{th}}$ electron to the dot. If an electron is added to a Bloch-state in 1D, 2D, or 3D, this electron can delocalize over the crystal and we can continue adding electrons as allowed by the DOS without an energy penalty. The only restriction is the Pauli exclusion principle, which forbids more than two electrons/allowed energy state. However, for a 0D structure, an added electron is spatially located close to the last electron. The electron-electron Coulomb repulsion raises the energy

[14] For example, the relation $n_s = \frac{C_b C_q}{C_b + C_q}(V - V_T)/q$ indicates that for $V < V_T$ the value of n is negative, which is impossible because n is an electron *density*, which can only be positive.

needed: this phenomena is called **Coulomb blockade**, which manifests in single-electron transistors, quantum-dot LEDs, and lasers, and in nanocrystal memories. The self-capacitance then is $C = q^2/\Delta E$, where ΔE is the difference in the eigenvalues that accounts for the electron-electron repulsion. This phenomena is similar to donors and acceptors, illustrated and discussed in Chapter 15, Fig. 15.6.

17.4 Kinetic inductance

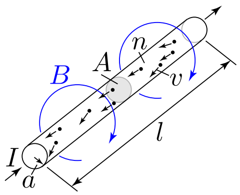


Fig. 17.14 External inductance, and kinetic inductance of a current-carrying wire surrounded by a medium of permeability μ.

In the last section, we found that a portion of energy of an electrostatic field was stored by charges inside the conductor, and used it to define the quantum capacitance. In a current-carrying conductor, additional energy is stored in the **kinetic energy** of the moving charges. The portion of energy stored in a magnetic field outside a conductor is its classical inductance, whereas the portion stored in the kinetic energy of moving electrons inside it is an extra inductance, called the **kinetic inductance**. Consider the conductor in Fig. 17.14 carrying a current I: it fills the space around it with a magnetic field $B = \mu I/2\pi\rho$ a distance ρ from the axis that curls around it. Because the energy stored per unit volume in a magnetic field is $\frac{B^2}{2\mu}$, the magnetic energy stored in the cylindrical shell $a < \rho < b$ (this is typically the case for coaxial cables) of length l due to the current flowing in the wire is

$$U = \frac{1}{2}LI^2 = l \cdot \int_{\rho=a}^{b} \frac{B^2}{2\mu} \cdot 2\pi\rho \cdot d\rho = \underbrace{\frac{\mu l}{4\pi}\ln[\frac{b}{a}]}_{\frac{1}{2}L} \cdot I^2 \implies L = \frac{\mu l}{2\pi}\ln[\frac{b}{a}], \tag{17.21}$$

where we note that the magnetic energy is proportional to the square of the electrical current via $U = \frac{1}{2}LI^2$, defining the inductance L. The physical meaning of inductance is that if we change the current I, the entire magnetic field energy has to respond, and that requires time and energy: this reluctance to change is captured by

$$V = -L\frac{dI}{dt}. \tag{17.22}$$

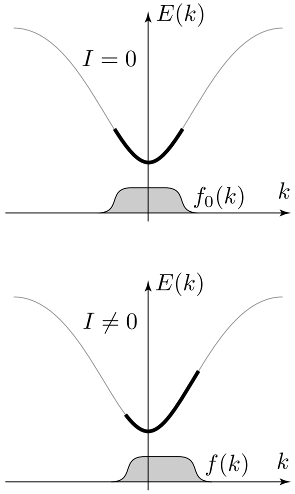
In the classical Drude picture, the extra energy or potential V necessary to cause this change is traced to the energy needed to change the magnetic dipole moments of the surrounding medium[15]. The inductance as understood above is the standard, or *magnetic* part. However, there is another *non-magnetic* part of the inductance that is typically neglected in conductors. Consider the wire in Fig. 17.14 of length l and a cross section area A, with mobile electron density n. A current $I = qnvA$ is carried by the mobile electrons moving at an ensemble velocity v. The total kinetic energy of the electrons is hence

$$\frac{1}{2}L_k I^2 = \text{K. E.} = \frac{1}{2}Mv^2 = \frac{1}{2}(nm_e lA)(\frac{I}{qnA})^2 \implies \boxed{L_k = \frac{l}{A}\frac{m_e}{q^2 n}.} \tag{17.23}$$

[15] The relative permeability μ_r is a measure of the magnetic polarizability with respect to vacuum. The vacuum, of course, has no atoms to polarize.

The kinetic energy defines another part of inductance, L_k, called the **kinetic inductance**, which depends on the electron mass and inversely on the electron density. Just like the external, magnetic part of the energy, there is an inertial resistance of the electrons to changing their density, velocity, and direction – the kinetic inductance is a measure of this inertia to change. The inductance increases for a larger (effective) mass, and increases for a low carrier concentration. A heavier object has more kinetic inertia, and for a low carrier concentration the carriers must move faster to carry the same current. A heavier object moving faster has a larger kinetic inductance. In typical metals the magnetic part of the inductance is dominant. But at high frequencies, and for very low resistance conductors, the kinetic inductance is not negligible[16].

Now let us find the kinetic inductance in the quantum picture. Before going into the details, we can guess that the quantum version will likely look like $L_k \sim \frac{m^\star}{q^2 n}$ where $m^\star$ is the band-effective mass, and n is the mobile carrier density, and this is indeed the form we will end up with. What follows is a careful bookkeeping of all states, and considering the complete Fermi–Dirac statistics. Consider a semiconductor (or any crystal: metal, semimetal, etc.) as the current carrier. The electrons carrying the current occupy the Bloch states of the bandstructure $E(\mathbf{k})$ with a certain occupation function. When the occupation function is the Fermi–Dirac distribution $f_0(\mathbf{k})$, the net current is zero because of exact cancellation of opposite currents (Fig. 17.15). When a current I flows, the bandstructure remains the same, but the occupation function must change to $f(\mathbf{k})$ to reflect this change.



$$L_k = 2 \cdot \frac{[U(I \neq 0) - U(I = 0)]}{I^2}$$

Fig. 17.15 Kinetic inductance is the effective inductor that leads to the net kinetic energy stored in the electron distribution carrying a current I.

There is an energy cost associated with this change: the net energy of electrons when the current flows is larger. The quantum inductance can be thus written in direct analogy to the classical kinetic inductance by first identifying the energy and current (the index $\sigma = \uparrow$ or $\downarrow$ is for spin):

$$U = \sum_{\mathbf{k},\sigma} E(\mathbf{k})[f(\mathbf{k}) - f_0(\mathbf{k})], \text{ and } I = \frac{q}{l} \sum_{\mathbf{k},\sigma} v_g(\mathbf{k}) f(\mathbf{k}), \tag{17.24}$$

where we have written the generalized expressions for electron transport in d-dimensions. Here l is the length (distance from source to drain contacts). Writing the kinetic energy as $U = \frac{1}{2} L_k I^2$ leads to a generalized kinetic inductance[17] for electron flow in d-dimensions:

$$\boxed{L_k = \frac{2U}{I^2} = \frac{2 \cdot \sum_{\mathbf{k},\sigma} E(\mathbf{k})[f(\mathbf{k}) - f_0(\mathbf{k})]}{(\frac{q}{l})^2 \cdot [\sum_{\mathbf{k},\sigma} v_g(\mathbf{k}) f(\mathbf{k})]^2} = \frac{l}{q^2 n_d} \cdot \frac{2[\frac{\sum_{\mathbf{k},\sigma} E(\mathbf{k})[f(\mathbf{k}) - f_0(\mathbf{k})]}{\sum_{\mathbf{k},\sigma} f(\mathbf{k})}]}{[\frac{\sum_{\mathbf{k},\sigma} v_g(\mathbf{k}) f(\mathbf{k})}{\sum_{\mathbf{k},\sigma} f(\mathbf{k})}]^2},} \tag{17.25}$$

where the form on the right is written as an analogy to the classical kinetic inductance. The numerator is an average energy, and the denominator is an average squared velocity: their ratio is an effective mass. Here, $n_d = \frac{1}{l \cdot W^{d-1}} \sum_{\mathbf{k},\sigma} f(\mathbf{k})$ is the volume density of electrons, and W^{d-1} is a cross-section: W^{d-1} is 1 for $d = 1$, a width W for $d = 2$, and a cross sectional area $W^2 = A$ for $d = 3$.

---

[16] The kinetic inductance plays a central role in **superconducting** wires and circuits, where $q \to 2q$ for Cooper pair carriers, and the carriers flow like a rigid correlated object. If scattering is reduced in semiconductors, then the kinetic inductance *increases* in importance as it is more difficult for electron velocities to change.

[17] This formulation may be used to evaluate the kinetic inductance for any material for which the bandstructure $E(k)$ is known, be it semiconductors, semimetals, topological insulators, etc. The inductance can be found for both the ballistic limit, or the scattering dominated limit. Also, see Exercise 17.4 to learn an equivalent formulation of the kinetic inductance $L_k = \frac{h}{q^2} \langle \frac{1}{2Mv} \rangle$ in terms of the modes M.

Parabolic Bandstructure: Let us find the kinetic inductance for a conductor of d-dimensions if its bandstructure is parabolic with an isotropic effective mass $E(k) = E_c + \frac{\hbar^2}{2m_c^\star}(\sum_{i=1}^{d} k_i^2)$. If the Fermi level is at E_F, the net carrier density is

$$n_d = N_c^d \cdot F_{\frac{d-2}{2}}(\frac{E_F - E_c}{k_b T})$$
$$= \frac{1}{2} N_c^d [F_{\frac{d-2}{2}}(\frac{E_{Fs} - E_c}{k_b T}) + F_{\frac{d-2}{2}}(\frac{(E_{Fs} - qV) - E_c}{k_b T})], \tag{17.26}$$

where E_{Fs} and E_{Fd} are the Fermi levels of the source and the drain contacts across which a voltage $E_{Fs} - E_{Fd} = qV$ is applied. The ballistic current density J_d in response to a voltage V applied to ohmic contacts at the source and drain is

$$J_d = \underbrace{\frac{q^2}{h} \cdot N_c^{d-1} \cdot \frac{k_b T}{q}}_{J_0} \cdot [F_{\frac{d-1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) - F_{\frac{d-1}{2}}(\frac{(E_{Fs} - qV) - E_c}{k_b T})]. \tag{17.27}$$

The net kinetic energy for a length L and cross-section W^{d-1} is

$$U(I = 0) = \underbrace{L \cdot W^{d-1} \cdot N_c^d \cdot d \cdot k_b T}_{U_0} \cdot F_{\frac{d}{2}}(\eta_F)$$
$$\implies U(I \neq 0) = U_0 \cdot [\frac{F_{\frac{d}{2}}(\eta_s) + F_{\frac{d}{2}}(\eta_s - v_d)}{2}]$$
$$\implies \text{K. E.} = U(I) - U(I = 0) = U_0[\frac{F_{\frac{d}{2}}(\eta_s) + F_{\frac{d}{2}}(\eta_s - v_d)}{2} - F_{\frac{d}{2}}(\eta_F)]. \tag{17.28}$$

And we get a general formulation for the quantum inductance for d-dimensions:

$$L_k = \frac{U_0}{(W^{d-1} J_0)^2} \cdot \frac{1}{2} \frac{[F_{\frac{d}{2}}(\eta_s) + F_{\frac{d}{2}}(\eta_s - v_d) - 2F_{\frac{d}{2}}(\eta_F)]}{[F_{\frac{d-1}{2}}(\eta_s) - F_{\frac{d-1}{2}}(\eta_s - v_d)]^2}$$
$$= \frac{L \cdot W^{d-1} \cdot N_c^d \cdot d \cdot k_b T}{(W^{d-1} \cdot \frac{q^2}{h} \cdot N_c^{d-1} \cdot \frac{k_b T}{q})^2} \cdot \frac{1}{2} \frac{[F_{\frac{d}{2}}(\eta_s) + F_{\frac{d}{2}}(\eta_s - v_d) - 2F_{\frac{d}{2}}(\eta_F)]}{[F_{\frac{d-1}{2}}(\eta_s) - F_{\frac{d-1}{2}}(\eta_s - v_d)]^2}$$
$$= \frac{L}{W^{d-1}} \times$$
$$\frac{m_c^\star}{q^2 \underbrace{N_c^d F_{\frac{d-2}{2}}(\eta_F)}_{n_d}} \cdot \underbrace{\left(\pi d F_{\frac{d-2}{2}}(\eta_F) \cdot \frac{[F_{\frac{d}{2}}(\eta_s) + F_{\frac{d}{2}}(\eta_s - v_d) - 2F_{\frac{d}{2}}(\eta_F)]}{[F_{\frac{d-1}{2}}(\eta_s) - F_{\frac{d-1}{2}}(\eta_s - v_d)]^2} \right)}_{\gamma_d}$$
$$\implies \boxed{\mathcal{L}_k = \gamma_d \cdot \frac{m_c^\star}{q^2 n_d}} = L_k \cdot \frac{W^{d-1}}{L}. \tag{17.29}$$

Table 17.1 Classical and quantum conductance, inductance, and capacitance.

Quantity	Bandstructure	1D	2D	3D
R_{cl}		$\frac{m^\star}{n_{1d}q^2\tau}\cdot l$	$\frac{m^\star}{n_{2d}q^2\tau}\cdot\frac{l}{W}$	$\frac{m^\star}{n_{3d}q^2\tau}\cdot\frac{l}{A}$
R_q	$E(k)=\hbar v_F k$	$\frac{h}{g_s g_v q^2}$	–	–
R_q	$E(k)=\frac{\hbar^2k^2}{2m^\star}$	$\frac{h}{g_s g_v q^2}$	–	–
L_{cl}		$\frac{\mu_0}{2\pi}\cdot\ln\left(\frac{2h}{d}\right)\cdot l$	–	–
L_k	$E(k)=\hbar v_F k$	$\frac{h}{2g_s g_v q^2 v_F}\cdot l$	$\frac{h}{2q^2 v_F k_F}\cdot\frac{1}{W}$	$\rho_{1d}\cdot l$
L_k	$E(k)=\frac{\hbar^2k^2}{2m^\star}$	$\frac{m^\star}{q^2 n_{1d}}\cdot l$	$\frac{\pi^2}{8}\cdot\frac{m^\star}{q^2 n_{2d}}\cdot\frac{l}{W}$	$\frac{4}{3}\cdot\frac{m^\star}{q^2 n_{3d}}\cdot\frac{l}{A}$
C_{cl}		$\frac{2\pi\epsilon_0}{\ln\left(\frac{2h}{d}\right)}\cdot l$	$\frac{\epsilon_0}{t_b}\cdot A$	$4\pi\epsilon_0 R_0$
C_q	$E(k)=\hbar v_F k$	$\frac{2g_s g_v q^2}{h v_F}\cdot\frac{1}{1+e^{-\frac{\mu}{k_bT}}}\cdot l$	–	–
C_q	$E(k)=\frac{\hbar^2k^2}{2m^\star}$	$q^2\cdot\frac{g_s g_v}{2\pi}\left(\frac{2m^\star}{\hbar^2}\right)^{\frac{1}{2}}\frac{1}{\sqrt{E_F-E_c}}\cdot l$	$q^2\cdot\frac{g_s g_v m^\star}{2\pi\hbar^2}\cdot\frac{1}{1+e^{-\left(\frac{\mu-E_c}{k_bT}\right)}}\cdot A$	$q^2\cdot\frac{g_s g_v}{(2\pi)^2}\left(\frac{2m^\star}{\hbar^2}\right)^{\frac{3}{2}}\sqrt{E_F-E_c}\cdot V$

The dimensional factor γ_d can be evaluated analytically in the degenerate limit at low temperatures and high carrier concentrations when $\eta_F >> +1$. The resulting values, and the corresponding quantum inductances are:

$$\begin{aligned} &\text{1D, } d=1:\gamma_1=1, \boxed{L_k^{1d}\approx L\cdot\frac{m_c^\star}{q^2n_{1d}}} \\ &\text{2D, } d=2:\gamma_2=\frac{\pi^2}{8}, \boxed{L_k^{2d}\approx\frac{L}{W}\cdot\frac{\pi^2}{8}\cdot\frac{m_c^\star}{q^2n_{2d}}} \\ &\text{3D, } d=3:\gamma_3=\frac{4}{3}, \boxed{L_k^{3d}\approx\frac{L}{W^2}\cdot\frac{4}{3}\cdot\frac{m_c^\star}{q^2n_{3d}}.} \end{aligned} \tag{17.30}$$

The quantum inductance per unit "length" for parabolic bandstructure condutors therefore is proportional to $m^\star/q^2n_d$ up to a factor of unity for all dimensions in the ballistic limit[18]. Table 17.1 shows a few examples of the classical and quantum $R, L,$ and C values. In this table, the respective conduction band-edge density of states for parabolic bandstrcuture is $g_d^p(E)=\frac{g_sg_v}{(4\pi)^{\frac{d}{2}}\Gamma(\frac{d}{2})}\left(\frac{2m_c^\star}{\hbar^2}\right)^{\frac{d}{2}}(E-E_c)^{\frac{d-2}{2}}$, and for conical bandstructure is $g_d^c(E)=\frac{g_sg_v}{(\hbar v_F)^d}\frac{1}{2^{d-1}\pi^{\frac{d}{2}}\Gamma(\frac{d}{2})}E^{d-1}$. The corresponding valence band-edge density of states are written down by analogy. There are a few missing spots in the Table: you can follow Exercise 17.6 to fill in these spots.

---

[18] Identical formulae are obtained for both the quantum capacitance and the kinetic inductance if the problem is solved in the scattering dominated regime using the Boltzmann transport equation of Chapter 21.

17.5 Quantum R, L, C circuits

Let us consider a metallic 1D carbon nanotube to gauge the importance of quantum capacitance and kinetic inductance. Fig. 17.16 shows a schematic figure of a "classical" conducting wire as a transmission line, connected to an AC voltage source, similar to the classical RLC circuit of Section 17.1 earlier in this chapter. The classical wave impedance is indicated as $Z_0 = \sqrt{\mu_0/\epsilon_0} \approx 120\pi = 377\ \Omega$. Also shown is a ballistic metallic carbon nanotube for comparison, which has a classical capacitance and inductance per unit length of roughly $C \approx 50$ aF/μm and $L \approx 1$ pH/μm. The metallic carbon nanotube has a linear energy band dispersion $E(k) = \hbar v_F k$ with a spin degeneracy $g_s = 2$ and valley degeneracy $g_v = 2$, and a characteristic band velocity $v_F \sim 8 \times 10^7$ cm/s. The linear dispersion implies the group velocity of each k-state is $v(k) = \frac{1}{\hbar}\frac{dE(k)}{dk} = v_F$. Since the DOS is $= \frac{2g_s g_v}{hv_F}$, the quantum capacitance per unit length is given by $C_q = q^2 \cdot \text{DOS} = g_s g_v \frac{q^2}{h}\langle\frac{2M}{v}\rangle = \frac{8q^2}{hv_F}$, which evaluates to $C_q \approx 386$ aF/μm. The kinetic inductance similarly is $L_k = \frac{h}{g_s g_v q^2}\langle\frac{2}{Mv}\rangle = \frac{h}{2q^2 v_F}$, which evaluates to $L_k \approx 16$ nH/μm. This implies the internal transmission line "quantum" wave impedance at high frequencies is $Z_q = \sqrt{\frac{L_k}{C_q}} = \frac{1}{G_q} = R_q = \frac{q}{g_s g_v h^2} \approx 6475\ \Omega$. The complex impedances dominate as $f \sim$ THz, highlighting a mismatch between the wave impedance in the ballistic nanotube with that of electromagnetic waves in vacuum[19]. The quantum capacitance and kinetic inductance of the metallic nanotube are therefore not negligible. It is important to account for the quantum capacitance in electrostatics, and both the quantum capacitance and kinetic inductance at high frequencies.

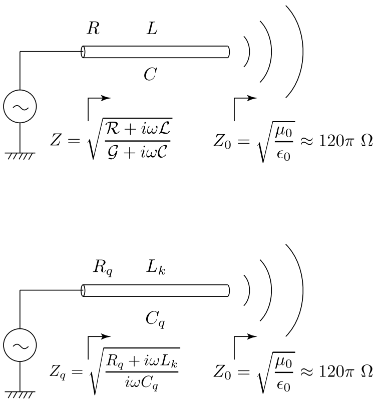

R
L
C
$Z = \sqrt{\dfrac{\mathcal{R} + i\omega\mathcal{L}}{\mathcal{G} + i\omega\mathcal{C}}}$
$Z_0 = \sqrt{\dfrac{\mu_0}{\epsilon_0}} \approx 120\pi\ \Omega$
R_q
L_k
C_q
$Z_q = \sqrt{\dfrac{R_q + i\omega L_k}{i\omega C_q}}$
$Z_0 = \sqrt{\dfrac{\mu_0}{\epsilon_0}} \approx 120\pi\ \Omega$

Fig. 17.16 Quantum limit of a RLC transmission line and its interface with the electromagnetic spectrum.

In Section 17.2, we argued that for long conductors that are many electron mean free paths $L >> \lambda_{mfp}$, one can neglect the role of the quantized conductance. Let us identify the situations when the quantum capacitance and kinetic inductance effects may be neglected. From the mode-picture (see Exercises 17.3 and 17.4), we can write the 1D per unit length quantities $C_q \sim \frac{MG_q}{\langle v\rangle}$ and $L_k \sim \frac{1}{MG_q\langle v\rangle}$. Since the 1D classical capacitance per unit length is $C_{cl} \approx 2\pi\epsilon_0\epsilon_r$, and the total capacitance is $C_{tot}^{-1} = C_{cl}^{-1} + C_q^{-1}$, we can neglect C_q if it is much *larger* than C_{cl}, which is achieved if

$$\frac{C_q}{C_{cl}} = \frac{2q^2 M}{hv} \cdot \frac{1}{2\pi\epsilon_0\epsilon_r} >> 1 \implies M >> \underbrace{\frac{4\pi\epsilon_0\hbar c}{q^2}}_{\frac{1}{\alpha}\sim 137} \cdot \frac{\langle v\rangle}{c} \cdot \frac{\pi\epsilon_r}{2} \sim \epsilon_r. \tag{17.31}$$

Here c is the speed of light, and $\alpha = \frac{1}{137}$ is the fine structure constant. The above analysis provides a simple rule of thumb: if the number of modes in the conductor exceeds roughly the relative dielectric constant of the medium surrounding the conductor, then the quantum capacitance may be neglected for the 1D conductor. For $\epsilon_r \sim 4$, roughly 10s of modes will make the classical capacitance dominate. The cases for 2D or 3D conductors should be evaluated separately since the electrostatic capacitance depends on the geometry.

---

[19] If the nanotube was to be used as a nanoscale antenna to radiate the energy as electromagnetic waves into the surrounding, the impedance mismatch is a measure of how difficult it is for it to radiate (or, because of reciprocity, detect).

Since the classical and kinetic inductances add, L_k should be much *smaller* than L_{cl} to be negligible. For the kinetic inductance to be negligible, a far higher number of modes is necessary. This is because the classical 1D capacitance and inductance per unit length of a transmission line are related via $L_{cl}C_{cl} = \frac{1}{c^2}$, whereas $C_q L_k \approx \frac{1}{\langle v \rangle^2}$, implying

$$\frac{L_{cl}C_{cl}}{L_k C_q} \sim \frac{\langle v \rangle^2}{c^2} \sim 10^{-4} \implies \frac{L_k}{L_{cl}} > 1 \text{ if } \frac{C_q}{C_{cl}} > 10^4. \tag{17.32}$$

From the discussion of the quantum capacitance, this implies a very large number of modes, exceeding thousands are necessary. In other words, the kinetic inductance is not negligible for nanoscale conductors and neglecting it would lead to significant errors in understanding and designing electronic devices that operate at high frequencies. The frequency range of interest can also be gauged by comparing the impedance due to kinetic inductance ωL_k with the resistance R. Using this limit, we can write (note that $R = L_k \cdot \tau$):

$$[\omega L_k = 2\pi f \cdot \frac{m^\star}{q^2 n} \cdot \frac{l}{A}] > [R = \frac{m^\star}{q^2 n \tau} \cdot \frac{l}{A}] \implies f > \frac{1}{2\pi\tau}, \tag{17.33}$$

where $\tau \sim 10^{-12}$ s is a typical scattering time. This analysis indicates that the kinetic inductance is important for high frequencies, beyond several 100s of GHz.

17.6 Negative R, L, and C

We have discussed the quantum limits of R, L, and C in the last three sections. An interesting question is: can the R, L, and C be *negative*? If the conductor is not *generating* energy, the answer is no[20]. Because if any of these three quantities were negative, from $dU/dt = I^2R$, $U = \frac{1}{2}LI^2$, and $U = \frac{1}{2}CV^2$, one can generate energy out of nowhere, violating energy conservation. On the other hand, the time-dependent, or complex impedance of each of these three elements in response to a driving current $I_\omega e^{i\omega t}$ or voltage $V_\omega e^{i\omega t}$ are $Z_R = R$, $Z_L = i\omega L$, and $Z_C = \frac{1}{i\omega C}$, where ω is the frequency. The complex portions of the impedances are not negative, but *imaginary* with a transparent physical meaning: the voltage drop across an inductor and a capacitor is out of phase with the current by either $+\pi/2$ or $-\pi/2$, whereas it is in phase with the current across a resistor. Though devices do not have negative DC resistance, inductance, and capacitance, they can have **negative *differential*** DC values of these three parameters. The negative differential values are not of purely academic interest, but have immense practical applications. For example, though a standard (positive) resistor cannot amplify electrical signals, if a device exhibits

---

[20] For example a solar cell under illumination *generates* energy, and the I-V curve goes into the fourth quadrant, making $I \cdot V$ negative (see Chapter 28). A simple $R = V/I$ will indeed give a negative resistance indicating that energy is being generated. Similarly, a thermoelectric device with a temperature gradient across it can generate energy from heat, as can a capacitor with metal plates that undergo radioactive decay that harvests nuclear energy. We do not consider these situations here.

NDR, it can be used in a feedback circuit to amplify or boost electrical signals.

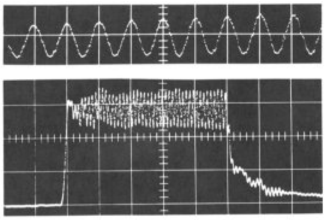


Fig. 17.17 Current-voltage oscillations observed by J. B. Gunn in n-type GaAs (from Sol. St. Comm., 1, 88, 1963).

Negative Differential Resistance (NDR): A strange property of the semiconductor GaAs was uncovered in 1962 by J. B. Gunn. He observed that when a lightly n-type doped semiconductor GaAs was contacted by heavily doped regions, and a DC current was passed through it, at certain DC currents the voltage drop across the device started *oscillating* with time, as seen in Fig. 17.17. Now a DC current $I = qnvA$, where n is the 3D electron density, $v = \mu F$ is the drift velocity in response to an electric field F with electron mobility μ ordinarily should lead to a single value of a DC voltage across a standard resistor.

The reason for the oscillatory output was soon tracked down by several researchers to the existence of a **negative differential resistance**, or NDR, originating in the **electron bandstructure** of GaAs. Fig. 17.18 shows the full bandstructure of GaAs calculated by the empirical pseudopotential method in Chapter 13. An interesting feature of the GaAs bandstructure is that the L-valley minimum is very close in energy to the Γ-valley minimum: the separation of the two valley minima is $E_L - E_\Gamma = E_{L-\Gamma} = 0.31$ eV, much smaller than the bandgap $E_g = 1.42$ eV. Also note that the L-valley minimum has a much smaller curvature than the Γ-valley minimum: indeed, the effective mass of the Γ valley minimum is $m^\star_\Gamma = 0.067 m_e$, whereas that of the L-valley is $m^\star_L = 0.55 m_e$, i.e., roughly 8X heavier than the Γ-valley. Finally, since the L-valley occurs at the BZ edge in the 111 direction, there are 8 such equivalent valleys, of which only half are in the 1st BZ, making the L-valley degeneracy $g_{vL} = 4$ compared to $g_{v\Gamma} = 1$.

The electrons at the bottom of the Γ-valley placed there by light donor doping are accelerated by an electric field via the standard rate equation $\mathbf{F} = \hbar \frac{d\mathbf{k}}{dt} - \frac{\mathbf{k}-\mathbf{k}_0}{\tau_\mathbf{k}}$. Since $m^\star_\Gamma$ is low, the change in $E(\mathbf{k})$ is rapid, and with a relatively small electric field of ~ 3 kV/cm, a significant population of electrons end up in the L-valley as shown schematically in Fig. 17.18. Because the effective mass in the L-valley is much larger, the electrons there have a much smaller drift velocity. The two carrier populations result in a *decrease* in the net velocity with an increase in the electric field, causing the negative differential resistance.

A simple model can capture this physics; the bottom panel of Fig. 17.18 shows a calculated velocity-field curve for GaAs. Assuming a lattice temperature of T_L, an electric field F causes the electrons to go out of equilibrium from the lattice, to an effective electron temperature T_e. This non-equilibrium condition is captured by the name **hot electrons**. The electron concentration in the Γ-valley in the non-degenerate situation is $n_\Gamma = N^{3d}_{c\Gamma} \cdot \exp{[-(E_\Gamma - E_F)/k_b T_e]}$, and in the L-valley is $n_L = N^{3d}_{cL} \cdot \exp{[-(E_L - E_F)/k_b T_e]}$. The ratio of electron densities in the L-valley to Γ-valley is

$$\frac{n_L}{n_\Gamma} = \frac{N^{3d}_{cL}}{N^{3d}_{c\Gamma}} \cdot \frac{e^{-\frac{E_L - E_F}{k_b T_e}}}{e^{-\frac{E_\Gamma - E_F}{k_b T_e}}} = \frac{g_{vL}}{g_{v\Gamma}} \cdot (\frac{m^\star_L}{m^\star_\Gamma})^{\frac{3}{2}} \cdot e^{-\frac{E_L - E_\Gamma}{k_b T_e}}. \tag{17.34}$$

By energy conservation, the rate of energy gain by electrons from the field must equal the rate at which the electrons lose their kinetic energy. The electrons lose their kinetic energy by colliding with the atoms of the crystal, setting off lattice vibrations, or phonons. This physics will be discussed in Chapter 22. Let the energy relaxation time be τ_E, which is typically of the order 10^{-12} s, or 1 ps. The total kinetic energy of electrons moving in 3D is $E = \frac{3}{2}k_b T_e$ for the hot electrons, and $E_0 = \frac{3}{2}k_b T_L$ when they are in equilibrium at $F = 0$. Then, the energy balance equation dictates that at steady state, we must have

$$\frac{dE}{dt} = qFv - \frac{E - E_0}{\tau_E} \underbrace{\implies}_{\text{steady state}} qFv = \frac{3}{2} \cdot \frac{k_b(T_e - T_L)}{\tau_E}$$

$$\implies T_e = T_L + \frac{2}{3} \cdot \frac{q\tau_E}{k_b} vF. \qquad (17.35)$$

The net current density due to electrons in both valleys is

$$J = q(n_\Gamma v_\Gamma + n_L v_L) = q(n_\Gamma + n_L)v \implies$$

$$v = \frac{n_\Gamma v_\Gamma + n_L v_L}{n_\Gamma + n_L} \underbrace{\approx}_{v_L \approx 0} \frac{\mu_\Gamma F}{1 + \frac{n_L}{n_\Gamma}} = \frac{\mu_\Gamma F}{1 + \frac{g_{vL}}{g_{v\Gamma}} \cdot (\frac{m_L^\star}{m_\Gamma^\star})^{\frac{3}{2}} \cdot e^{-\frac{E_L - E_\Gamma}{k_b T_e}}}, \qquad (17.36)$$

where the ensemble velocity of electrons is expressed as a function of the bandstructure parameters, the driving electric field F, and the resulting electron temperature T_e. We have made the reasonable approximation that the velocity of electrons in the L-valley is far smaller than in the Γ-valley, and neglected their contribution to the current, but retained their contribution to the net electron concentration. The electron temperature is related to the electric field F and the lattice temperature T_L and the bandstructure via

$$T_e = T_L + \frac{2}{3} \cdot \frac{q\tau_E \mu_\Gamma F^2}{k_b[1 + \frac{g_{vL}}{g_{v\Gamma}} \cdot (\frac{m_L^\star}{m_\Gamma^\star})^{\frac{3}{2}} \cdot e^{-\frac{E_L - E_\Gamma}{k_b T_e}}]}. \qquad (17.37)$$

This equation is numerically solved to obtain the self-consistent electron temperature T_e for a given electric field F and lattice temperature T_L. Then, the velocity-field curve is obtained for a T_L by varying the electric field using Equation 17.36, using an experimental electron mobility $\mu_\Gamma \sim 9400 \cdot (300/T_L)$ cm^2/V·s where T_L is the lattice temperature in Kelvin. The bottom panel of Fig. 17.18 shows the resulting electron velocity as a function of the electric field, and the fraction $n_L/(n_\Gamma + n_L)$ of the electrons that are in the L-valley. For fields less than ~ 2 kV/cm, the velocity increases sharply, but because of the transfer of electrons to the high effective mass L-valley, the net velocity decreases for electric fields higher than $F \sim 3$ kV/cm. As a result, the **current decreases with increasing voltage: this is the NDR**. Cooling the device enhances the NDR. Note that there is no negative resistance, only negative *differential* resistance: V/I is always positive unless the device is generating energy.

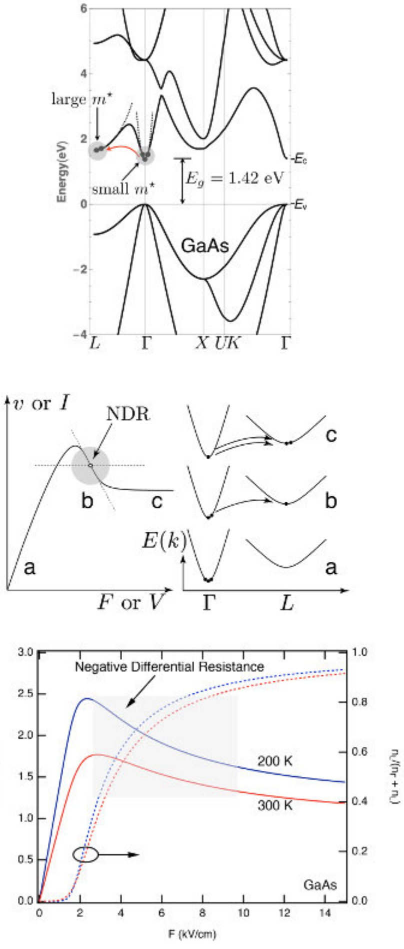


Fig. 17.18 Physics of the negative differential resistance (NDR) observed in the Gunn effect, and its relation to the bandstructure of GaAs. The transfer of electrons from a high velocity valley to a low velocity valley (or the *transferred electron* effect) is responsible for the NDR. Note that the NDR is reached at a relatively low electric field of ~ 3 kV/cm, which makes Gunn oscillators useful for compact and effective sources of microwave power.

If a fixed DC current is driven in the NDR regime, there are two stable voltage values. Local fluctuations in the carrier concentration set up oscillations between the allowed voltage points. Thus, the simple device is able to convert a portion of the DC power $I \cdot V$ supplied to it into AC power $\Delta I \cdot \Delta V$, where ΔI and ΔV are the margins of oscillations. The frequency of these oscillations are limited by the $\Gamma \to L$ intervalley scattering rate, to be discussed in Chapter 23. Typical frequencies achieved are in the 1–100 GHz range, making the Gunn oscillator an effective microwave power source[21]. The two-terminal device is connected to metal antennas of appropriate size to radiate and receive microwave signals. Gunn oscillations are typically observed in bulk semiconductors that have bandstructure traits similar to GaAs (such as InP, etc.), and is a beautiful example of the immense potential hidden in the details of the $E(k)$ diagram of crystals, stemming from the quantum mechanics of electrons in periodic potentials. NDR in the current-voltage characteristics is also obtained in heavily doped p-n junction diodes (due to Esaki tunneling) – to be covered in Chapters 24, and in heterostructure resonant tunneling diodes, also be covered in Chapter 24. But the simplicity of construction and operation of the Gunn oscillator, which needs just a bulk semiconductor is the reason for its enduring popularity.

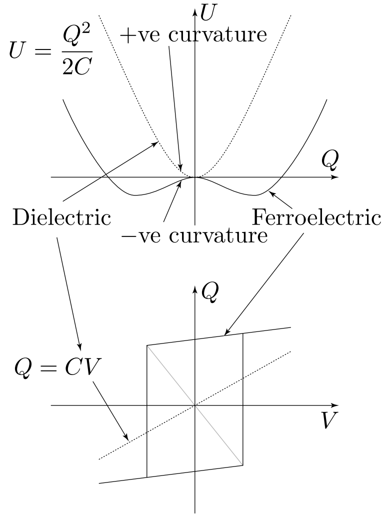


Fig. 17.19 Negative Differential Capacitance.

Negative Differential Capacitance and Inductance: Unlike the resistor which dissipates energy, a capacitor C and an inductor L can *store* energy. When electron spins interact strongly with each other and switch collectively, the solid can become ferromagnetic. Ferromagnetic materials are used in inductor cores to boost their energy storage capacity. Similarly, when electric dipoles interact strongly with each other and switch collectively, the material becomes ferroelectric. Both ferromagnetic and ferroelectric materials are characterized by a response function that exhibits hysteresis. For a ferromagnet, the net magnetization is hysteretic with the external magnetic field, whereas for a ferroelectric material the net electrical polarization, or charge Q is hysteretic with the external voltage V, or the electric field. Fig. 17.19 indicates a ferroelectric energy vs. charge curve for a parallel plate ferroelectric capacitor, and the corresponding hysteretic loop of charge in response to the voltage, compared to a standard (non-ferroelectric) dielectric. Because the ferroelectric has two lowest energy states, poled "up" and "down" (similar to N and S of a ferromagnet), the minima of energy U vs. Q has a hump, with a negative curvature region around the origin. This is somewhat similar to the effective mass in electronic bandstructure: therefore, it is believed that a ferroelectric capacitor can potentially provide a DC negative differential capacitance. Similarly, a DC negative differential inductance can also be expected in an inductor. These properties are currently under investigation. Since some semiconductors can be ferroelectric, and some can also be ferromagnetic, negative differential capacitance and inductance effects in DC could prove to be a useful for new generations of device applications to add to negative differential resistance[22].

[21] Because the Gunn oscillator is a compact source of microwave power at high frequencies ($\sim$ 1–100 GHz), and it produces the microwave power at a low voltage input with respectable conversion efficiency, it is used in radar detectors, collision avoidance radar, movement sensors, and several related applications.

[22] Because of their hysteresis in their properties, both ferromagnets and ferroelectrics are used for memory elements.

The concepts of quantized conductance, capacitance, and inductance are important in the field of **circuit quantum electrodynamics**, which provides the conceptual framework for the design of quantum bits or **qubits**, the workhorse for quantum computation. In such circuits, electromagnetic energy is manipulated in the form of a single photon.

17.7 Chapter summary section

In this chapter, we learned:

- How to view conductance, capacitance, and inductance of conductors from the point of view of their bandstructure and electronic modes.
- The mechanism of ballistic transport, and how to move smoothly between the ballistic regime to the scattering dominated regimes using the powerful concept of quasi-Fermi levels.
- The concepts of quantum capacitance and kinetic inductance, which assume increasing importance in nanoscale conductors, and at high frequencies.
- How negative differential resistance can be achieved in bulk semiconductors exploiting quirks of the electron bandstructure.

Further reading

Quantum Transport: Atom to Transistor, and **Lecture Notes in Nanoelectronics** by S. Datta provide lucid introductions to the concepts, and many facets of quantized conductance and quantum capacitance, and are highly recommended as further reading. Several concepts discussed in this chapter are topics of current research and I expect them to undergo refinements with new experiments. Aspects of kinetic inductance and quantum capacitance that are currently unknown are discussed in the review article **Nanoelectromagnetics: Circuit and Electromagnetic Properties of Carbon Nanotubes** by Rutherglen and Burke, in Small (2009), volume 5, page 884.

Exercises

(17.1) **Quantum current flow and saturation in semiconductors**

We discussed the carrier density in semiconductor bands and the ballistic current flowing in them for transport in $d = 1, 2, 3$ dimensions in Chapter 5.

(a) Consider the 1D case, encountered in semiconductor quantum wires and carbon nanotubes. Use a conduction band effective mass of $m_c^\star = 0.2m_e$, valley degeneracy $g_v = 1$ and spin degeneracy $g_s = 2$. Calculate and plot the source quasi-Fermi level at the source injection point $(E_{Fs} - E_c)$, and the drain quasi Fermi level $(E_{Fs} - qV - E_c)$ as a function of the voltage $0 < V < 2$ Volt for a high electron density $n_{1d} = 5 \times 10^6/\text{cm}$ at room temperature $T = 300$ K, and explain the plot. Next, plot the ballistic currents vs. voltage for electron densities $n_{1d} = 1, 2, 3, 4, 5 \times 10^6/\text{cm}$, both at 300 K and at 77 K for $0 < V < 2$ Volt. You should observe that in 1D semiconductors the ballistic current does *not* depend on the 1D density at low voltages and low temperatures – why is this so? Why does the current saturate at high voltages?

(b) Now consider the 2D case, which is encountered in silicon transistors, III-V quantum well high-electron mobility transistors, and 2D crystal semiconductors and metals. For this problem, use a conduction band effective mass of $m_c^\star = 0.2m_e$, valley degeneracy $g_v = 1$ and spin degeneracy $g_s = 2$. Calculate and plot the source quasi-Fermi level at the source injection point $(E_{Fs} - E_c)$, and the drain quasi Fermi level $(E_{Fs} - qV - E_c)$ as a function of the voltage $0 < V < 2$ Volt for a high electron density $n_{2d} = 5 \times 10^{13}/\text{cm}^2$ at room temperature $T = 300$ K, and explain the plot. Next, plot the ballistic current per unit width vs. voltage for electron densities $n_{1d} = 1, 2, 3, 4, 5 \times 10^{13}/\text{cm}^2$, both at 300 K and at 77 K for $0 < V < 2$ Volt. You should observe that unlike in 1D semiconductors, the ballistic current per unit width in 2D *does* depend on the 2D density at low voltages and low temperatures – why is this so? Why does the current saturate at high voltages?

In this problem, you have solved the 1D and 2D ballistic transistor problem in disguise. The variation of the carrier density in a transistor is done with a third terminal called the gate. We will use the same method to understand the behavior of the ballistic field-effect transistor (FET) in Chapter 19.

(17.2) **A 2DEG as a parallel array of 1D conductors**
Electrons of sheet carrier density n_s sit in the conduction band of a 2D electron system of energy bandstructure $E(k_x, k_y) = \frac{\hbar^2}{2m_c^\star}(k_x^2 + k_y^2)$ with the k-space occupation of carriers shown in Chapter 5, Fig. 5.14. Assume a spin degeneracy of $g_s = 2$ and a valley degeneracy of $g_v = 1$. The width of the 2D system is W, the length L, and ohmic source and drain contacts are made to connect to the electrons to flow a current in the x-direction. Solve this problem entirely at $T = 0$ K. The allowed discrete points in the k-space $(k_x, k_y) = (\frac{2\pi}{L}n_x, \frac{2\pi}{W}n_y)$ where (n_x, n_y) are integers are are considered individual *modes* of the 2DEG as indicated in Fig. **??**. The collection of modes with the same n_y is considered a 1D mode of the 2DEG.

(a) When the applied voltage across the source/drain contacts is $V_{ds} = 0$, find the Fermi wavevector k_0 as shown in the left of Fig. 5.14.

(b) Show that the number of 1D modes with current flow in the x-direction because of the finite width of the 2D conductor is $M_0 = \frac{k_0 W}{\pi}$. Use part (a) to write this in terms of the 2DEG density.

(c) Now a voltage V_{ds} is applied across the drain and the source such that the net sheet carrier density of the 2DEG does not change. Assume ballistic transport and show that in Fig. 5.14, $k_R = \sqrt{k_0^2 + \frac{m_c^\star}{\hbar^2}(qV_{ds})}$ and $k_L = \sqrt{k_0^2 - \frac{m_c^\star}{\hbar^2}(qV_{ds})}$.

(d) Show that the voltage V_{ds} reduces the total number of left going modes M_L and increases the total number of right going modes M_R. Find expressions for M_L and M_R.

(e) Find the voltage V_{ds} at which carriers in all modes move to the right and no carriers move to the left.

(f) Find how many right-going 1D modes are present in the above situation when all carriers move to the right.

(g) Because each 1D mode in the ballistic limit can provide the maximum conductance of a quantum of conductance $G = \frac{g_s g_v q^2}{h}$, find the "saturation" current I_d when the critical V_{ds} of part (e) is reached.

(17.3) **Quantum capacitance as a function of modes**
Show that the quantum capacitance derived in 17.11 may be equivalently reformulated as

$$C_q = \frac{q^2}{h}\langle\frac{2M}{v}\rangle, \tag{17.38}$$

where M is the mode number, v is the group velocity of state k, and the average $\langle f \rangle = \frac{\sum_k f}{\sum_k}$ is a sum over k-states.

(17.4) **Kinetic inductance as a function of modes**
Show that the quantum capacitance derived in

17.11 may be equivalently reformulated as

$$L_k = \frac{h}{q^2} \langle \frac{1}{2Mv} \rangle, \quad (17.39)$$

where M is the mode number, v is the group velocity of state k, and the average $\langle f \rangle = \frac{\sum_k f}{\sum_k}$ is a sum over k-states.

(17.5) **Kinetic inductance of graphene**
Show that the quantum inductance of 2D graphene of bandstructure $E(k_x, k_y) = \hbar v_F \sqrt{k_x^2 + k_y^2}$ at $T \to 0$ K is

$$L_k = \frac{h}{2q^2 v_F k_F} \cdot \frac{1}{W}, \quad (17.40)$$

where h is Planck's constant, k_F is the Fermi wavevector, defining the Fermi level $E_F = \hbar v_F k_F$, and W is the width.

(17.6) **Filling in the missing quantized R, L, and C**
Fill in the missing terms and complete Table 17.1 by analogy to the examples provided throughout this chapter. Also compare and contrast how the R_q, L_k, and C_q change with dimensions and bandstrutures.

Junction Magic: Schottky, pn and Bipolar Transistors

18

18.1 Ballistic Schottky diodes 401
18.2 pn diodes: discovery 408
18.3 pn diodes: transport 411
18.4 Bipolar junction transistors 417
18.5 Deathniums! 423
18.6 Chapter summary section 424
Further reading 426
Exercises 426

At the heart of every semiconductor device is a Schottky diode, or a pn junction diode in one form or another. A clear understanding of the physics of their operation is essential for the analysis, and design of modern electronic and photonic devices. The majestic range of applications of the Schottky and pn junction diodes arise from the physics of electron transport in them. Unlike the resistors, capacitors, and inductors of Chapter 17, controlling the transport of electrons or holes in diodes enables us to engineer highly non-linear (and even negative differential) resistors, and transistors leading to **electronic gain**. In Chapter 15, we discussed the equilibrium energy band diagrams of these diodes. In this chapter, we will understand:

- Ballistic electron transport in Schottky diodes and the resulting current-voltage characteristics,
- How the capacitive behavior of Schottky diodes are at the root of field-effect transistors,
- Electron and hole transport in pn junction diodes, and
- How the transport physics of the pn diode is responsible for electronic gain, bipolar transistors, solar cells, LEDs and lasers.

18.1 Ballistic Schottky diodes

Historically, the Schottky diode is the first semiconductor device in which *unidirectional*, or *non-reciprocal* current flow was observed. Metal-semiconductor rectifiers were investigated as early as mid 1880s by Braun (Fig. 18.1), and used by Bose (Fig. 18.2) to demonstrate usage in short wavelength (high frequency) radio receivers, followed by Marconi's entrepreneurship that gave birth to radio communications. All this happened well before the physics of metals and semiconductors were understood through band theory! But the understanding of the physics of the device after the revelations of the quantum mechanical theory of metals and semiconductors as we have discussed in Modules I and II in this book, propelled them into a class of very useful electronic devices via several breakthroughs. Initially the unreliability of the cat's whiskers was removed, soon leading to power rectifiers, transistors, and much more.

Fig. 18.1 Ferdinand Braun in 1875 used metal-semiconductor point contact rectifiers. For his development of phased array antennas, he was awarded the 1909 Nobel Prize in Physics, with Marconi. Note that the electron was yet to be discovered, and semiconductors with the associated quantum mechanics and solid state physics were half a century away.

Fig. 18.2 Jagadish Bose demonstrated the use of metal-semiconductor contacts for the earliest mm-wave radio receivers in 1890s.

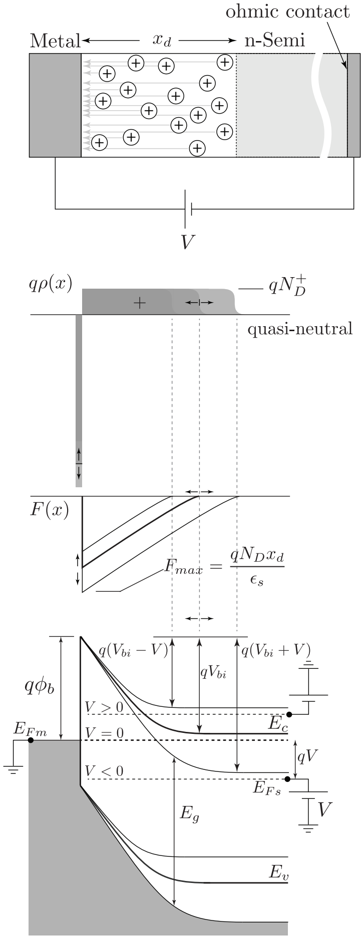


Fig. 18.3 Biased Schottky diode: charge-electric field-energy band diagrams.

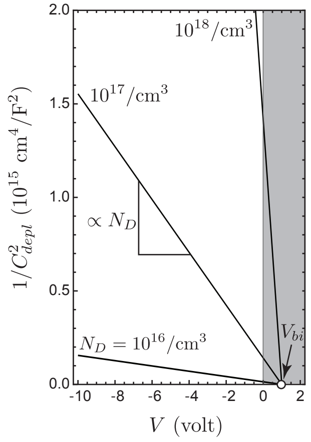


Fig. 18.4 A plot of $1/C^2_{depl}$ vs. voltage directly yields N_D and V_{bi}.

In Chapter 15, Fig. 15.11 we have discussed the equilibrium energy band diagram and electrostatics of the Schottky diode. The major conclusions were: due to the built-in voltage from the difference in the work functions $qV_{bi} = E_{Fm} - E_{Fs}$ at equilibrium, charge transfer occurs between the metal and the semiconductor. For a rectifying contact, the n-type semiconductor donor-doped at a density N_D loses mobile carriers to the metal, forming a depletion region of thickness $x_d = \sqrt{2\epsilon_s V_{bi}/qN_D}$ at the junction. There is a potential barrier formed for electrons to move from the conduction band of the semiconductor into the metal due to the depletion region. That is our point of departure for discussing how the application of a voltage bias leads to a rectifying current. Let us for a moment neglect the flow of current, and focus on how the depletion region, and the barrier height changes upon applying a voltage bias across the Schottky diode. This is shown in Fig. 18.3. A battery of voltage V is now connected across the contacts. For clarity of the following discussion, we will assume that the metal is grounded, and the battery is connected to the semiconductor through an ohmic contact far from the junction, as indicated in the energy band diagram of Fig. 18.3. This means the Fermi level of the semiconductor E_{Fs} is controlled by the voltage of the battery. The equilibrium case is when $V = 0$, and $E_{Fm} = E_{Fs}$.

Reverse Bias: Application of a -ve voltage $V < 0$ on the terminal (called *reverse* bias) connected to the metal is equivalent to applying a +ve voltage at the terminal connected to the semiconductor, if the metal were grounded. This must charge the semiconductor more positive than it is at equilibrium. Since $p << n$ in the n-doped semiconductor, there are very few mobile positive charges. So how exactly does this happen? Note that the dominant positive charges in this Schottky diode are the immobile ionized donors, of volume density N_D^+. Removal of mobile electrons will increase the positive charge in the semiconductor. So when a negative voltage V is applied on the battery, it extracts a slice of mobile electrons from the edge of the depletion region at $x = x_d$, changing the depletion region thickness to

$$x_d' = \sqrt{\frac{2\epsilon_s(V_{bi}-V)}{qN_D}} \implies C_{depl} = \frac{\epsilon_s}{x_d'} = \sqrt{\frac{qN_D\epsilon_s}{2(V_{bi}-V)}}, \tag{18.1}$$

and the corresponding depletion region capacitance C_{depl} per unit area changes. This effect is used to directly measure V_{bi} as the intercept and N_D as the inverse slope in a plot of $1/C^2_{depl}$ vs. the applied voltage V (Fig. 18.4). Where do the extracted electrons go when a reverse bias voltage V is applied? The battery pumps these electrons into the metal, and they end up in the sheet charge within a lattice constant on the metal side of the M-S junction as indicated in Fig. 18.3. Note that the system on the whole is charge neutral, as it always should be, and the battery is also charge neutral. Importantly, the electrons are transferred from the semiconductor to the metal not over the potential barrier in the conduction band (because the barrier height has *increased*), but via the *external circuit*, through the battery and the wires. If this voltage is maintained on the battery, the effective barrier height for electrons at the flat part of the conduction band edge in the semiconductor for $x > x_d$ to go to the semiconductor has *increased* from $qV_{bi} \to q(V_{bi} - V)$ since V is negative. The strength of the electric field in response to the voltage also increases. Note that since the *slope* of the electric field by Gauss's law is $\frac{dF(x)}{dx} = -\frac{qN_D^+}{\epsilon_s}$, it does not change in shape inside the depletion region if the doping density is uniform. Thus the electric field profile $F(x)$ in Fig. 18.3 changes as similar triangles. The change in the area under the electric field curve $\int F(x)dx$ is simply the applied voltage V, leading to a surface field

$$F_{max} = \frac{qN_D x_d'}{\epsilon_s} = \sqrt{\frac{2qN_D(V_{bi} - V)}{\epsilon_s}}. \tag{18.2}$$

Forward Bias: Now if a +ve voltage $V > 0$ is applied on the battery, E_{Fs} must increase, and the semiconductor should get charged more negatively than in equilibrium. The battery achieves this by pulling some electrons from the metal through the external circuit and pumping them into the semiconductor through the ohmic contact. The expressions for the depletion region thickness, the depletion capacitance, and the maximum surface electric field F_{max} in Equations 18.1 and 18.2 remain unchanged for $V > 0$. This means that the depletion region shrinks, and the surface electric field decreases at $V > 0$ than the equilibrium. Of high importance is the lowering of the potential barrier for electrons at the conduction band edge of the semiconductor near the depletion edge to $q(V_{bi} - V)$, which will now lead to an exponential increase in the current. Before significant current flows however, the above discussions show that the Schottky diode is a variable capacitor (or varactor) in its low-conductivity state.

Fig. 18.5 shows the calculated depletion thickness, depletion capacitance, and the surface electric field for three different doping densities. The depletion thickness is smallest for the heaviest doping. The depletion capacitance decreases at negative (reverse bias) voltages as $1/\sqrt{V_{bi} - V}$, and the corresponding surface electric field increases as the reverse bias voltage increases. For example, for a heavy doping of $N_D = 10^{18}/\text{cm}^3$, the depletion thickness is $x_{depl} \sim 40$ nm, the zero bias depletion capacitance is C_{depl} =0.27 μF/cm^2, and the surface electric field is $F_{max} = 0.6$ MV/cm. We will see shortly that a high surface electric field under reverse bias promotes the quantum mechanical tunneling of electrons through the bandgap. But first let us look at current flow in the absence of tunneling.

Fig. 18.6 (a) shows a detailed energy-band diagram of the Schottky diode when a forward bias voltage V is applied. The quasi Fermi level of the bulk n-type semiconductor is raised by qV with respect to the metal Fermi level: $E_{Fs} - E_{Fm} = qV$. Note that the top of the barrier is shown as rounded compared to the sharp kind shown in gray, and discussed earlier. This is an important real effect, which we

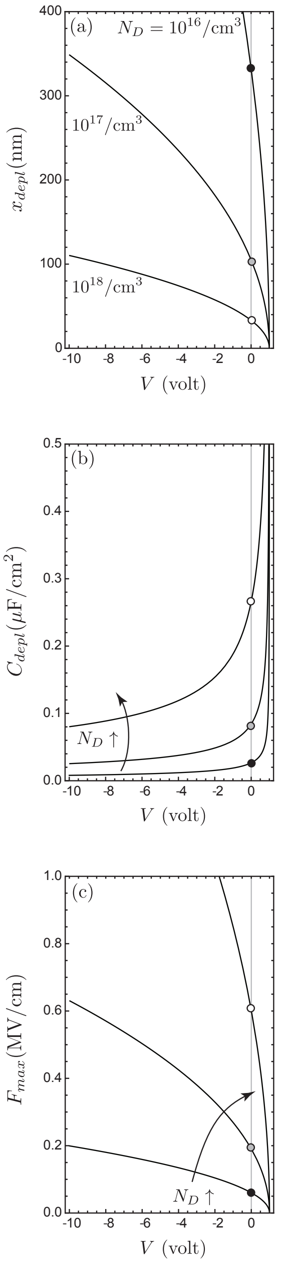


Fig. 18.5 Schottky diode depletion thickness x_d, depletion capacitance C_{depl}, and surface electric field F_{max} as a function of voltage V for $N_D = 10^{16}, 10^{17}$, and $10^{17}/\text{cm}^3$ for a semiconductor of dielectric constant $\epsilon_s = 10\epsilon_0$ and a M-S built-in voltage $V_{bi} = 1$ V.

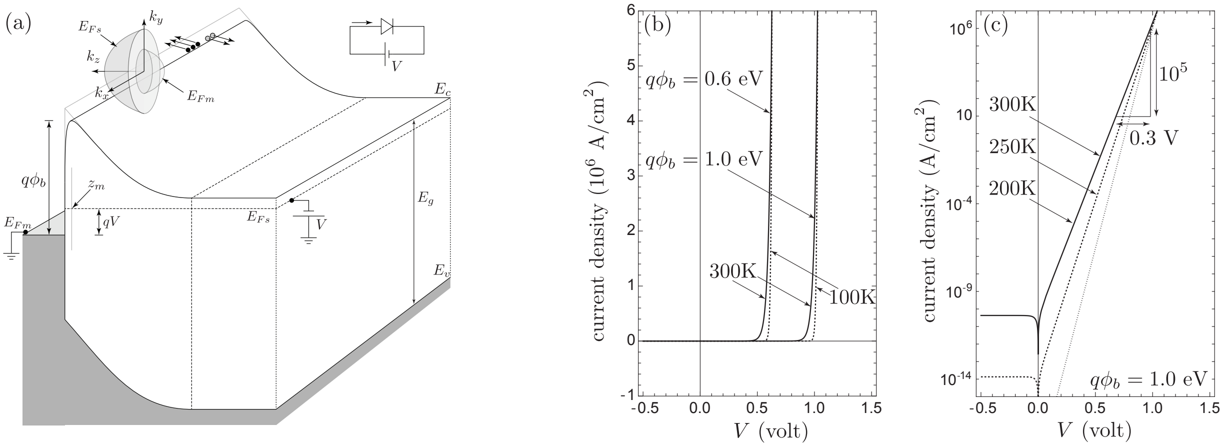


Fig. 18.6 (a) Energy band diagram of a Schottky diode showing the image-force lowering of the barrier height, the electron distribution in **k**-space at the top of the barrier at forward bias, and the circuit diagram. (b) Linear scale current-voltage characteristics of Schottky diodes of two barrier heights $q\phi_b = 0.6$ V and $q\phi_b = 1.0$ V at 300 K and 100 K. (c) Log scale current-voltage characteristics of a Schottky diode of barrier height $q\phi_b = 1.0$ V at three different temperatures. The calculations are for an isotropic, single-valley ($g_v = 1$) conduction band minimum with an electron effective mass $m_c^\star = 0.2m_e$.

now discuss to prepare for calculating the current.

The band edge variation from the junction $z = 0$ without image force lowering (Fig. 18.7) varies quadratically from Poisson equation:

$$E_c(z) = q(\phi_b - V_{bi} - V) + \frac{q^2 N_D (z - z_d)^2}{2\epsilon_s}, \text{ where } V_{bi} = \frac{qN_D z_d^2}{2\epsilon_s}, \tag{18.3}$$

where z_d is the depletion thickness. Consider an electron a distance z from the junction, in the depletion region of the semiconductor, approaching the top of the barrier. The free–electrons in the metal rearrange to create an image charge of the electron of positive sign located at a point $-z$ inside the metal. As a result, the electron experiences an attraction $V_{img}(z) = -q^2/4\pi\epsilon_s(2z)$. Adding this image potential to $E_c(z)$ shifts the barrier maximum from the junction to a point slightly away from the junction inside the semiconductor at z_m. The top of the barrier is lowered by a finite amount, and the new barrier height is what we re-label as $q\phi_b$ as also shown in Fig. 18.6 (a)[1].

We now evaluate the current in the full ballistic approximation by neglecting tunneling. The quantum mechanical current density is given, as always, by $J = \frac{qg_sg_v}{L^3}\sum_{\mathbf{k}} v_g(k) \cdot f(k) \cdot T(k)$ where g_s, g_v are the spin and valley degeneracies, $v_g(k)$ is the group velocity of state **k** and $f(k)$ its occupation function, and $T(k)$ is the transmission probability. A very useful trick to calculate currents is to realize the obvious: due to current continuity, **the current density J is the same at all planes** z. Consider the plane $z = z_m$ in Fig. 18.6 (a) where the potential barrier reaches its maximum after image force lowering. In this plane, the electrons in the conduction band moving to the left towards the metal must have come from the bulk of the semi-

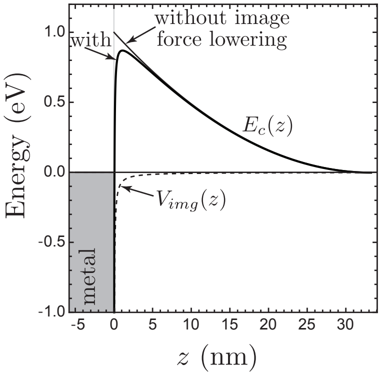


Fig. 18.7 Image force lowering of the Schottky barrier height.

---

[1] The top of the barrier plane is very close to the junction ($z_m/z_d << 1$), and the lowering of the barrier height is also very small compared to the total barrier height for a high dielectric constant and standard doping densities. Try out Exercise 18.2 for strange things that may happen in non-standard cases.

conductor. Therefore the left-moving electron population is in thermal equilibrium with the semiconductor bulk, meaning it shares the quasi Fermi level E_{Fs} of the semiconductor bulk. The occupation function of left going electrons is then the Fermi–Dirac function $f_L(E) = (1+\exp[(E-E_{Fs})/k_bT])^{-1}$. This is shown schematically as a hemisphere in the **k**-space where the highest occupied states have energy E_{Fs}. In the same plane, the electrons moving to the right into the semiconductor bulk must have come from the metal, and thus share the quasi Fermi level E_{Fm} and have the corresponding occupation function $f_R(E) = (1+\exp[(E-E_{Fm})/k_bT])^{-1}$. The two quasi Fermi levels are related by the applied voltage $E_{Fs} - E_{Fm} = qV$. Note from the energy band diagram of Fig. 18.6 (a) that at the top of the barrier plane $z = z_m$, $E_c - E_{Fs} = q(\phi_b - V)$ where V is the applied voltage, and $E_c - E_{Fm} = q\phi_b$. The total ballistic current density flowing across this interface is therefore[2]

$$J = J_L - J_R = \frac{q g_s g_v}{L^3}\Big[\sum_{\mathbf{k}\in\mathbf{k_L}} v_g(k) f_L(E) - \sum_{\mathbf{k}\in\mathbf{k_R}} v_g(k) f_R(E)\Big] =$$

$$\frac{q g_s g_v}{L^3}\Big[\int_{-\infty}^{+\infty}\frac{dk_x}{\frac{2\pi}{L}}\int_{-\infty}^{+\infty}\frac{dk_y}{\frac{2\pi}{L}}\int_{-\infty}^{0}\frac{dk_z}{\frac{2\pi}{L}}\cdot\frac{\hbar k_z}{m_c^\star}\cdot\frac{1}{1+\exp\big[\frac{E_c+\frac{\hbar^2}{2m_c^\star}(k_x^2+k_y^2+k_z^2)-E_{Fs}}{k_bT}\big]}$$

$$-\int_{-\infty}^{+\infty}\frac{dk_x}{\frac{2\pi}{L}}\int_{-\infty}^{+\infty}\frac{dk_y}{\frac{2\pi}{L}}\int_{0}^{+\infty}\frac{dk_z}{\frac{2\pi}{L}}\cdot\frac{\hbar k_z}{m_c^\star}\cdot\frac{1}{1+\exp\big[\frac{E_c+\frac{\hbar^2}{2m_c^\star}(k_x^2+k_y^2+k_z^2)-E_{Fm}}{k_bT}\big]}\Big]$$

$$= \underbrace{\frac{2\pi g_s g_v q m_c^\star k_b^2}{h^3}}_{A^\star\text{ = Richardson constant}}\cdot T^2\cdot\Big[F_1\Big(\frac{q(V-\phi_b)}{k_bT}\Big) - F_1\Big(-\frac{q\phi_b}{k_bT}\Big)\Big]$$

$$F_j(\eta << -1) \approx e^{\eta} \implies \boxed{J = J_L - J_R \approx \underbrace{A^\star\cdot T^2\cdot e^{-\frac{q\phi_b}{k_bT}}}_{J_0}\cdot[e^{\frac{qV}{k_bT}}-1]}$$

$$\boxed{J = J_0[e^{\frac{qV}{k_bT}}-1]}\ \textbf{The diode equation.} \quad (18.4)$$

In the evaluation above, we integrate over the left going $k_z < 0$ and right going $k_z > 0$ hemispheres separately, retaining the full Fermi–Dirac distributions. The exact result is expressed in the form of the Fermi–Dirac integral $F_1(\eta)$ of order 1 and the arguments $\eta_L = q(V-\phi_b)/k_bT$ for the left going electrons and $\eta_R = -q\phi_b/k_bT$ for the right going electrons. Since $k_bT \sim 0.026$ eV whereas $q\phi_b \sim 1.0$ eV, $\eta_R << -1$, we can safely make the Maxwell–Boltzmann approximation of the Fermi–Dirac integral. For the left going electrons, for $V < \phi_b$, the Maxwell–Boltzmann approximation is a good one too; in realistic Schottky diodes the current is so high for $V \approx \phi_b$ that $V_{ext} = I\cdot R_{ext}$ voltage drops external to the diode kick in and restrict the actual drop across the Schottky diode to $V < \phi_b$ for all practical cases. The calculated Schottky diode currents are shown in Figures 18.6 (b) and (c). They highlight several aspects of the nature of the diode equation $J = J_0(e^{qV/k_bT} - 1)$:

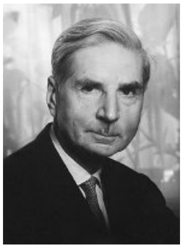


Fig. 18.8 Walter Schottky was an early investigator of the metal-semiconductor junction. A semiconductor research institute in Munich is named after him.

---

[2] We have considered that the transmission probability for an electron with energy larger than the barrier is 1. This is a fair assumption if the potential changes gradually, the quantum mechanical transmission coefficient then is indeed unity. Here we have considered a semiconductor with an isotropic conduction bandstructure $E_c + \frac{\hbar^2}{2m_c^\star}(k_x^2+k_y^2+k_z^2)$, and that there are g_v identical valleys with this same bandstructure. If the bandstructure is not isotropic, and the valleys are not identical, the method of evaluation remains identical, but we must account for the correct addition of the effective mass terms in the transport direction in each valley. See Exercise 18.3 for an example for multivalley anisotropic Schottky currents.

- The coefficient J_0 contains the **Richardson constant** $A^\star$. This is a historically important constant (see Fig 18.9) because it contains a combination of fundamental constants and the bandstructure of the semiconductor via the effective mass $m_c^\star$ and the valley degeneracy g_v. The plots in Figures 18.6 (b) and (c) are for $m_c^\star = 0.2 m_e$ and $g_v = 1$. The dependence on the bandstructure parameters is linear. The bandstructure dependence of the Richardson constant for anisotropic multivalley semiconductors is discussed in Exercise 18.3.
- The coefficient J_0 contains the strong, exponential temperature-dependent and barrier-height dependent term $T^2 e^{-q\phi_b/k_b T}$. Thus the diode current "turns on" around $V \approx \phi_b$. This is seen in the linear current-voltage plots of Fig. 18.6 (b). The voltage turn on is seen vividly around $V \approx \phi_b$ for the two barrier heights of 0.6 eV and 1.0 eV, the competing terms of $e^{-q\phi_b/k_b T}$ and $e^{+qV/k_b T}$ show that the voltage overcomes the potential barrier by raising the conduction band edge and injecting a large amount of electrons that move to the left, while the applied voltage has no effect on the right going electrons since for them the barrier height remains $q\phi_b$. The linear plot indicates that the effect of temperature is not that strong in the on-state current of the diodes.
- The voltage dependent term $(e^{qV/k_b T} - 1)$ is critical to why the diode is such a useful semiconductor device. The temperature-dependence is vividly seen in the "off-state" of the diode when plotting in the log-scale as shown in Fig. 18.6 (c). The reverse-bias current for $V < 0$ is J_0, which decreases exponentially with temperature. The current density increases exponentially also for voltages $V < \phi_b$. A characteristic feature of thermionic emission currents, and in general, for currents flowing over a potential barrier due to thermal injection from electrodes is the $e^{qV/k_b T}$ behavior: to change the current by a factor of 10 (a decade), a voltage $\ln 10 \times k_b T \sim 2.303 \times k_b T$ must be applied in the forward bias. This is referred to as the **inverse subthreshold slope (SS)** and is very important for diodes as well as transistors. For example, at room temperature since $k_b T \sim 26$ meV, the SS is ~ 60 mV/decade: in Fig. 18.6 (c) it is seen that at 300 K, an increase in voltage by 0.3 V = 5×60 mV results in an increase of the diode current by 10^5. Typical Schottky diode currents measured experimentally are indeed exponential well over 10 orders of magnitude, and the temperature dependence is used in experiments to extract various parameters of the diode.
- We evaluated the quantum current in the plane coincident with the top of the potential barrier, and neglected tunneling currents. To explore more about evaluating ballistic currents deep inside the semiconductor, and including tunneling currents, see Exercise 18.4.
- The circuit diagram of the Schottky diode is shown in Fig. 18.6

Fig. 18.9 Owen Richardson was awarded the 1928 Physics Nobel Prize for his studies of thermionic emission; the Richardson constant is named after him. For the emission of electrons from a metal into vacuum, the mass is that of a free–electron, making the original Richardson constant $A = \frac{4\pi q m_e k_b^2}{h^3}$, somewhat like the Lorenz number, a combination of fundamental classical and quantum constants q, m_e, k_b, and h. Because unlike emission of electrons from metal to vacuum, in a semiconductor the band parameters change $m_e \rightarrow m^\star$ and can have multiple valleys, the semiconductor Richardson coefficient changes from $A \rightarrow A^\star$, just like the mass $m^\star$, ionization energy of donors $Ry^\star$, effective Bohr radius $a_B^\star$, etc.

Fig. 18.10 Hans Bethe in his short report of war-time microwave detectors evaluated the Schottky diode current in the way closest to what we have done in this section. He identified it as a thermionic emission process: Mott and Schottky before him had used drift/diffusion methods to evaluate the current density. Bethe was a giant in every area of physics, ranging from atomic and solid-state physics to nuclear and astrophysics. He was awarded the 1967 Nobel Prize for discovering how stars generate energy. Freeman Dyson has called Bethe the "Supreme problem solver of the 20th century".

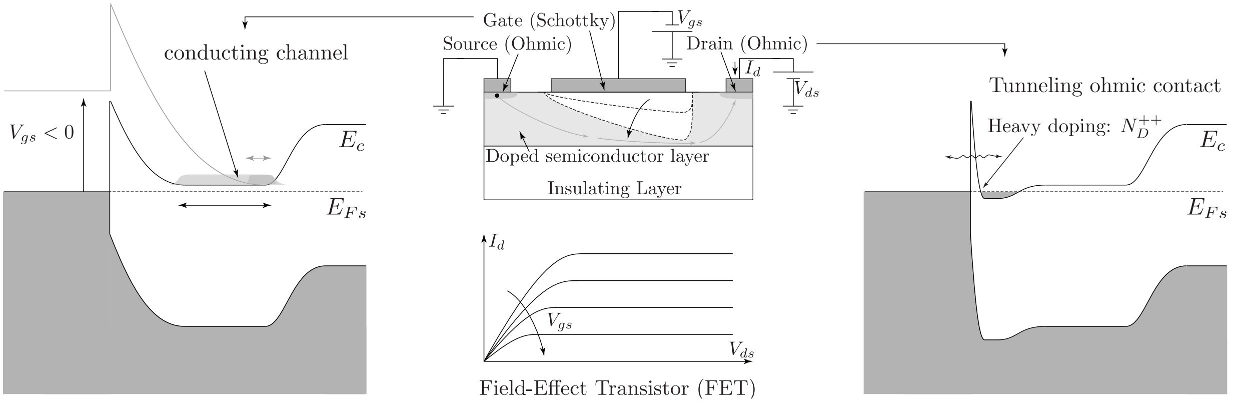


Fig. 18.11 Figure illustrating the construction of a field-effect transistor (FET) by the use of metal-semiconductor contacts of two types. The source and drain contacts are formed of heavily doped tunneling Schottky contacts, whereas the gate is formed of a near-ideal Schottky contact that modulates the current flow by depleting the thin conductive channel in the semiconductor located on an insulating semiconducting substrate.

(a). The rectifying action is evident in Fig. 18.6 (b): for $V < 0$ and also for $0 < V \leq \phi_b$, the diode blocks the current efficiently, but lets large amounts of current to flow for $V \approx \phi_b$ and higher. If an alternating current is made to flow through the Schottky diode, only the direction allowed will flow in each cycle, while the other half will be blocked. If a current at a frequency ω_0 is fed to the Schottky diode biased near $V \approx \phi_b$, the highly non-linear $J - V$ characteristics will produce several higher harmonics at frequencies $2\omega_0, 3\omega_0, ...$: in this way the Schottky diode is used as a mixer in electronic circuits.

Schottky Gate Transistor: Fig. 18.11 shows how the combination of Schottky junctions come together to form a metal-semiconductor field-effect transistor (MESFET). The semiconductor bulk is insulating (typically called semi insulating or S.I. bulk), except the top layer is doped n-type with shallow donors to provide mobile electrons in this layer. A Schottky gate metal is deposited on the surface, forming the energy band diagram shown on the left. The conducting channel is the region that is not depleted under the gate metal. When a negative gate voltage is applied on the gate, the conductive channel is depleted, increasing the resistance of the channel. Under the source and drain contacts, a thin layer of the semiconductor is heavily doped, forming the energy band diagram on the right. Because the depletion region is very thin here, electrons can easily tunnel through the thin barrier, forming the ohmic contacts. The current in response to a fixed drain voltage is therefore modulated by the gate voltage. Because the effective voltage drop across the gate-source is smaller than the gate-drain, the depletion region is thicker on the drain side. At a fixed drain voltage, the gate is able to remove the conducting channel, turning the device off. At a fixed gate voltage, the drain current increases with

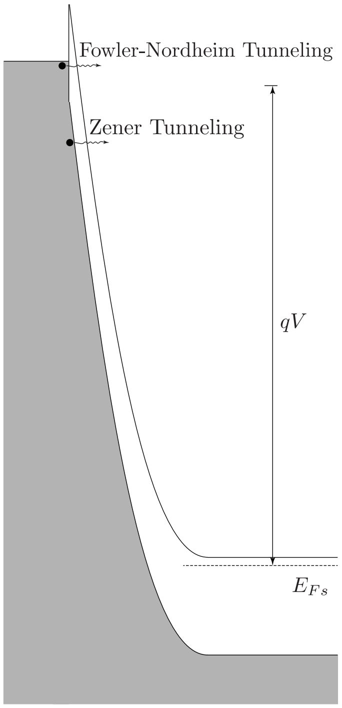


Fig. 18.12 At large reverse bias voltages, tunneling currents flow in the Schottky diode that add and exceed the standard thermionic part given by $A^{\star}T^2 e^{-q\phi_b/k_b T}$. Tunneling occurs both of electrons from the metal (the Fowler–Nordheim process), and those in the valence band of the semiconductor (the Zener process). These quantum tunneling processes are discussed further in Chapter 24.

the drain voltage, but reaches saturation when the depletion edge approaches the semi-insulating layer. We will discuss more about the working principles of the FET in **Chapter 19**.

Fig. 18.12 shows a Schottky junction at a very high reverse bias voltage. The electric field near the surface increases sharply at high reverse bias voltages, making it possible for electrons in the metal to tunnel through the triangular barrier into the semiconductor conduction band. Compared to exponentially small electrons that can be thermionically emitted over the barrier with near unity transmission probability, the tunneling current with a lower transmission probability wins out because of the enormous density of electrons below the metal Fermi level. Tunneling of electrons from the metal into the semiconductor is called the **Fowler–Nordheim tunneling** process, and sometimes referred to as **field-ionization**. Electrons from the valence band of the semiconductor can also tunnel into the conduction band to the right at the same energy, but for them the barrier height is the entire bandgap of the semiconductor, which is larger than the Schottky barrier height. This feature makes the interband tunneling mechanism (called **Zener tunneling**) contribute less than the Fowler–Nordheim tunneling when the two compete[3]. Both intra- and inter-band tunneling processes in semiconductors will be discussed in **Chapter 24** in more depth.

Fig. 18.13 In 1940, working at the Bell Laboratories, Russell Ohl discovered an enormous photovoltage develop across a cracked junction of a silicon crystal. This was simultaneously the first pn junction, and the first semiconductor photovoltaic cell. The physical mechanism leading to the operation of a pn junction diode was soon unraveled by Bell Labs scientists, which would lead to the discovery of the bipolar junction transistor, and semiconductor photodetectors, LEDs, and lasers.

18.2 pn diodes: discovery

In this chapter till now, we have discussed Schottky diodes formed by metal/n-type semiconductor junctions. Prior to the 1940s, n-type and p-type semiconductors were not clearly distinguishable, and the concept of doping itself was not understood. In Bell Laboratories, Scaff, Theurer, and Ohl (Fig. 18.13) observed that for certain silicon wafers, the Schottky diodes would turn on when a **p**ositive voltage was applied to the metal, whereas for some others it would turn on if a **n**egative voltage was applied to the metal (Fig. 18.14). They named it p-type Si and n-type Si to demarcate this difference in conductivity. Yet the microscopic origin of this property remained mysterious for a while. In 1940, Ohl discovered to his amazement that for a large silicon crystal, the nature of the semiconductor changed across a visible junction, producing a very large photovoltage across the junction when visible light was shining on it: this was the first ever pn junction diode, *and* the first **solar cell** to boot!

A serendipitous incident gave the first clue into the microscopic origin of doping of semiconductors, and led to the unraveling of the physics of the pn junction diode. Theurer noticed that when n-type silicon wafers were freshly cleaved, he could literally *smell* phosphorus in the air[4]. It was observed that whenever phosphorus could be detected on cleaving silicon, the corresponding Schottky diode would turn on at a positive voltage. The connection was made that phos-

---

[3] This feature is responsible for the observation that a pn junction diode can sustain a higher breakdown voltage than a Schottky diode. In high voltage pn diodes, the breakdown is typically not due to Zener tunneling, but due to avalanche multiplication.

[4] It turns out that some human noses are as sensitive to the smell of phosphorus as the best equipment!

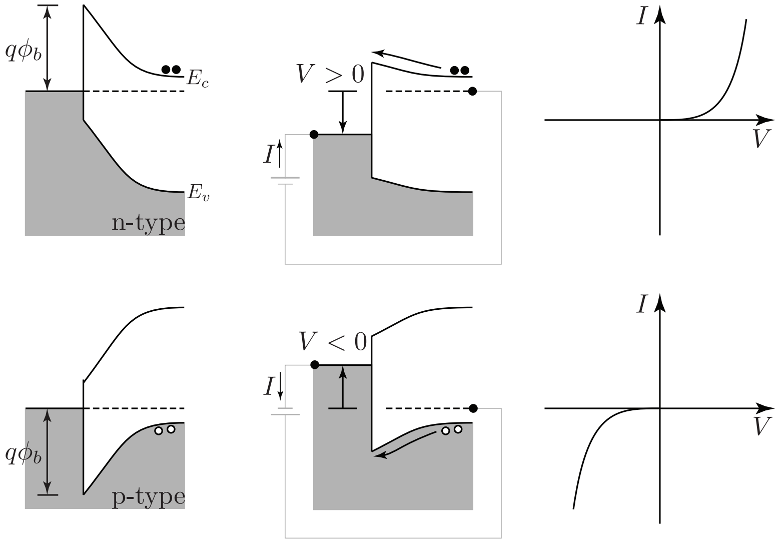


Fig. 18.14 Why metal/n-type semiconductor and metal/p-type semiconductor Schottky diodes turn on in opposite directions. Schottky barrier diodes are **majority-carrier devices**, which dictate the direction of current flow. Note that unlike in the metal/n-type junction, the Schottky barrier is for holes in the valence band for the p-type semiconductor.

phorus is on the right side of silicon in the periodic table with one extra electron. When it substitutes in a silicon crystal site, it **dopes the silicon crystal n-type**, putting an extra electron in the conduction band. The silicon crystals were initially grown in alumina crucibles at high temperatures. The outer crust of the silicon crystal, which was in contact with the crucible was observed to be p-type. Aluminum is just to the left of silicon in the periodic table. When aluminum substitutes a silicon atom, it removes an electron from the valence band of the semiconductor, creating a hole, and thereby **doping it p-type**. It was confirmed that Schottky diodes made on the outer crust of the silicon crystals turned on when a negative voltage was applied to the metal, proving the concept of **substitutional donor and acceptor doping** for the first time. This was the birth of chemical doping of semiconductors, paving the path to the pn junction diode.

In Chapter 15, Sections 15.4 and 15.5 we discussed the electrostatics and energy band diagrams of a pn junction diode at equilibrium, when the voltage bias is zero. Fig. 18.15 reproduces the equilibrium charge/field/band diagrams in light gray scale, and the solid darker lines show how these change when a forward bias voltage is applied across the junction. The built-in voltage $V_{bi} = (k_b T/q)\ln[N_A N_D/n_i^2]$ does not change with voltage, since it depends only on the doping densities N_D, N_A on the two sides of the junction and the semiconductor material through n_i. qV_{bi} is precisely the potential energy barrier for electrons in the conduction band on the n-side to flow to the conduction band on the p-side, as indicated in the first energy band diagram[5]. Application of a forward bias V lowers this barrier to $q(V_{bi} - V)$, as indicated. W_n and W_p are the thicknesses of the undepleted regions. The depletion region shrinks, and the triangular electric field profile reduces in strength such that the change in the area under the field is equal to the applied voltage V. The new relations are:

---

[5] Note that since $n_i^2 = N_c N_v e^{-E_g/k_b T}$, this built-in potential is $qV_{bi} = E_g - k_b T \ln[(N_c N_v)/(N_A N_D)]$, slightly smaller than the bandgap of the semiconductor.

$$V_{bi} - V = \frac{1}{2} F_{max}(x_n + x_p) \implies F_{max} = \sqrt{\frac{2qN_A N_D (V_{bi} - V)}{\epsilon_s (N_A + N_D)}},$$

$$x_n + x_p = W = \sqrt{\frac{2\epsilon_s}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right) \cdot (V_{bi} - V)}, \text{ and}$$

$$x_n = \frac{N_A}{N_A + N_D} W, \text{ and } x_p = \frac{N_D}{N_A + N_D} W. \quad (18.5)$$

The ratio of the depletion thicknesses x_n and x_p remain identical to the equilibrium condition to maintain charge neutrality. When forward biased, the battery increases the density of mobile electrons in the n-side and mobile holes in the p-side near the depletion region edge, shrinking the depletion region. This increase is shown in the logarithmic plot of the equilibrium carrier densities n_{n0}, p_{n0} on the n-side and p_{p0}, n_{p0} on the p-side. The increase in the mobile carrier concentrations inside, and near the edge of the depletion region are crucial to understand the flow of current in the diode.

At forward bias, electrons – the majority carriers on the n-side of density $n_{n0} \approx N_D$ are injected into the depletion region. Some of them overshoot and make it across the depletion region into the p-side. The increase in the electron density inside the depletion region, and in the p-region is indicated by the solid line in the logarithmic scale in Fig. 18.15. Electrons are **minority carriers** in the p-type region. Since there are a lot more holes in the p-side in the valence band of density $p_{p0} \approx N_A$, an injected electron will diffuse briefly in the p-side before recombining with a hole. This interband recombination may be radiative in direct-bandgap semiconductors **producing light**, or non-radiative in indirect bandgap semiconductors producing heat. The diffusion length for minority electrons is $L_n = \sqrt{D_n \tau_n}$, where D_n is the diffusion constant and τ_n the minority carrier lifetime as long as the undepleted p-region is much thicker than L_n. The excess minority electron density on the p-side decays away from the depletion region as e^{-x/L_n}, which we will derive in the next section. A mirror image of this picture occurs for holes: they are injected as minority carriers into the n-side, and diffuse there for a length $L_p = \sqrt{D_p \tau_p}$ till they recombine with electrons. The increase in the majority and minority mobile carrier concentrations at all x at applied voltage V is captured by the local quasi Fermi levels F_n of electrons, and F_p of holes via

$$n(x) = N_c e^{-\frac{E_c(x) - F_n(x)}{k_b T}} \text{ and } p(x) = N_v e^{-\frac{F_p(x) - E_v(x)}{k_b T}}$$

$$\implies \boxed{n(x) \cdot p(x) = n_i^2 \cdot e^{\frac{F_n(x) - F_p(x)}{k_b T}}} \text{ at all points in space.} \quad (18.6)$$

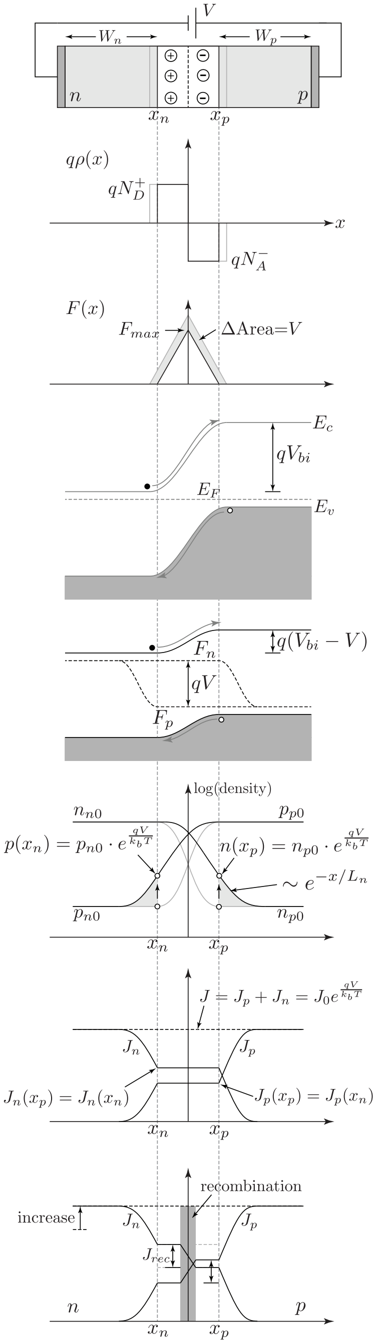


Fig. 18.15 The pn junction diode at forward bias. The charge, field, energy band diagrams, and electron and hole densities and currents are shown.

18.3 pn diodes: transport

The concept of quasi Fermi levels was introduced by Shockley (Fig. 18.16), who gave it a special name: **imref**, (i.e., Fermi spelt backwards, following Fermi's suggestion!), to highlight its central importance in much of semiconductor device physics (we will continue calling it the quasi Fermi level). We will now see how this concept is able to simplify the analysis of current flow in the pn junction diode. In Fig. 18.15, it is seen that when no voltage is applied, there is one Fermi level E_F. When a forward bias voltage V is applied across the junction, the Fermi level splits into the quasi Fermi levels F_n and F_p as shown in the junction. Analogous to the concept in ballistic transport, the mobile electrons in the conduction band are in equilibrium with the n-contact metal voltage in the undepleted n-side, and through the depletion region, as captured by a constant F_n. Since the density of the injected minority electrons decreases by recombination in the undepleted p-side, F_n drops in the p-side. The same is true for holes in the valence band: they are in equilibrium with the p-contact, sharing a quasi Fermi level F_p in the p-side, and through the depletion region, before recombination with electrons increases F_p in the undepleted n-side. Note that far away from the depletion edges, F_n and F_p merge into one Fermi level. These regions have the same electron and hole densities as they had in equilibrium. **The split in the quasi Fermi level is a measure of how much the carrier density product $n(x) \cdot p(x)$ deviates from its intrinsic, or thermal equilibrium value of n_i^2.** The maximum splitting inside the depletion region is equal to the applied voltage: $F_n - F_p = qV$.

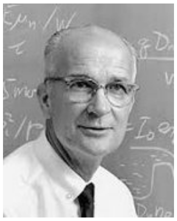


Fig. 18.16 William Shockley provided the theory of the operation of the pn junction, immortalized in the equation $J = q(n_p \frac{D_n}{L_n} + p_n \frac{D_p}{L_p})(e^{qV/k_bT} - 1)$. Shockley is one of the pioneers of semiconductor physics and devices, and was awarded the Nobel Prize in Physics in 1956 with Bardeen and Brattain for the invention of the transistor. His startup company "Shockley Semiconductor Laboratory" established the silicon valley, though he earned a reputation of being a difficult boss!

Consider the fate of a minority electron that has made it from the n-side, across the depletion region, into the p-side. Let $\Delta n(x) = n(x) - n_{p0}$ be the excess electron density above the equilibrium minority electron density n_p for $x > x_p$. We can write down the continuity equation from Equation 16.26:

$$\frac{\partial \Delta n(x)}{\partial t} = -\frac{d}{dx}\left(\frac{J_n}{-q}\right) + (G_n - R_n). \tag{18.7}$$

Inside the undepleted p-region, the electric field is very small[6], we approximate it to be negligible: $F(x) \to 0$. Since the total electron current J_n is the sum of the drift and diffusion components $J_n = qn\mu_n F(x) + qD_n \frac{dn(x)}{dx}$, and the recombination rate in the p-side is $R_n = \frac{\Delta n(x)}{\tau_n}$, the continuity equation becomes

$$\frac{\partial \Delta n(x)}{\partial t} = D_n \frac{d^2 \Delta n(x)}{dx^2} - \frac{\Delta n(x)}{\tau_n}. \tag{18.8}$$

At steady state, $\frac{\partial ...}{\partial t} = 0$, leading to the diffusion equation

$$\frac{d^2 \Delta n(x)}{dx^2} = \frac{\Delta n(x)}{D_n \tau_n} \implies \Delta n(x) = Ae^{+\frac{x}{L_n}} + Be^{-\frac{x}{L_n}}, \tag{18.9}$$

[6] This assumption is best justified *a posteriori*, after evaluating the current density. The current deep inside the undepleted regions is carried by drift where a very small field is needed because the majority carrier density is large. A comprehensive form of the current is to express it in terms of the quasi Fermi levels directly: $J_n = n\mu_n \frac{dF_n}{dx}$ and $J_p = p\mu_p \frac{dF_p}{dx}$. Using Equation 18.6, this leads to $J_n = qn\mu_n F + qD_n \frac{dn}{dx}$ and $J_p = qp\mu_p F - qD_p \frac{dp}{dx}$. From this formulation, it is seen that the quasi Fermi levels change very slowly when the current component is carried by majority carriers, and fast when it is carried by minority carriers.

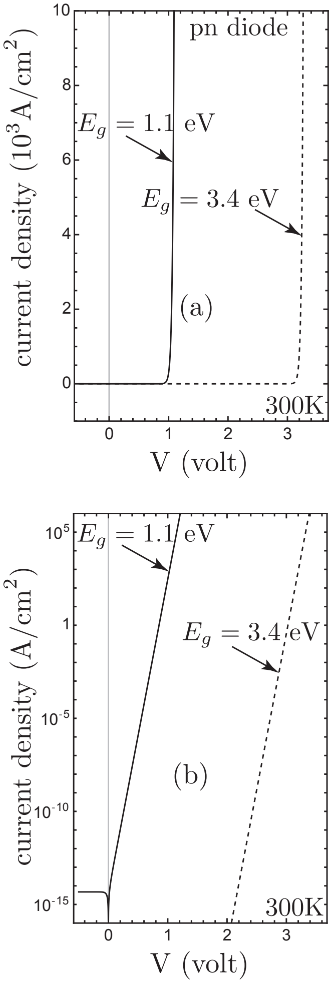


Fig. 18.17 Plots of the pn junction current $J = q(n_{p0}\frac{D_n}{L_n} + p_{n0}\frac{D_p}{L_p})(e^{qV/k_bT} - 1)$ for two semiconductors in linear and log scales. The doping densities are $N_A = N_D = 10^{17}/\text{cm}^3$, band edge effective masses $m_c^\star = m_v^\star \approx 0.2m_e$, electron and hole mobilities $\mu_n = 800$ cm^2/V·s and $\mu_p = 300$ cm^2/V·s for both semiconductors. The minority carrier lifetimes are $\tau_n = 30$ μs and $\tau_p = 10$ μs for the semiconductor of bandgap $E_g = 1.1$ eV, and $\tau_n = \tau_p = 1$ ns for $E_g = 3.4$ eV. The current densities after turn-on are in the ~ 10 kA/cm^2 range. The diffusion lengths $L_n = \sqrt{D_n\tau_n}$ for these examples are ~ 200 μm for the narrow bandgap semiconductor and ~ 1 μm for the wide bandgap semiconductor because of the different minority carrier lifetimes.

where $L_n = \sqrt{D_n\tau_n}$ emerges as the natural length scale over which the minority carrier concentration decays exponentially due to recombination. A and B are unknown coefficients to be determined from boundary conditions. As is seen in Fig. 18.15, far from the depletion edge, all *excess* minority carriers have recombined, leading to $\Delta n(x \rightarrow +\infty) = 0 \implies A = 0$. At the edge of the depletion region, $F_n - F_p = qV$. The minority carrier concentration is obtained by the Shockley boundary condition at $x = x_p$ using Equation 18.6 that $n(x_p)\cdot p(x_p) = n_i^2 e^{qV/k_bT}$, and since $p(x_p) \approx p_{p0} \approx N_A$, the minority density is increased exponentially from it's equilibrium value of n_{p0} to $n(x_p) \approx n_{p0}\cdot e^{qV/k_bT}$. Thus, setting $x = x_p = 0$, we get

$$\Delta n(x=0) = n_{p0}(e^{\frac{qV}{k_bT}} - 1) = B$$

$$\implies n(x) - n_{p0} = n_{p0}(e^{\frac{qV}{k_bT}} - 1)\cdot e^{-\frac{x}{L_n}}$$

$$\implies J_n^{diff}(x) = qD_n\frac{dn(x)}{dx}|_{x=0} = -qD_n\frac{n_{p0}}{L_n}(e^{\frac{qV}{k_bT}} - 1)\cdot e^{-\frac{x}{L_n}}$$

$$\implies \boxed{|J_n^{diff}(x_p)| = qD_n\frac{n_{p0}}{L_n}(e^{\frac{qV}{k_bT}} - 1),\ \text{Diffusion current.}} \qquad (18.10)$$

Note how the spatial variation of the excess minority carrier density directly leads to the diffusion current density, also shown in Fig. 18.15. The characteristic $e^{qV/k_bT} - 1$ dependence on the applied voltage makes its appearance. The absolute value indicates the magnitude of the current that flows towards -ve x, which we skip in further discussion since the direction of flow of currents is qualitative clear. Now Shockley's simplifying insight was to consider the case when the net electron (or hole) current density remains constant across the depletion region, i.e., $J_n(x_n) = J_n(x_p)$. The total current, the sum of the electron and hole currents $J_{tot} = J_n(x) + J_p(x)$ is constant at all x to ensure the continuity of the total current. Therefore, we must have the following magnitude of the diode current:

$$J_n(x_n) = J_n^{diff}(x_p) \implies J_{tot} = J_n^{diff}(x_p) + J_p^{diff}(x_n)$$

$$\boxed{|J_{tot}| = q(n_{p0}\cdot\frac{D_n}{L_n} + p_{n0}\cdot\frac{D_p}{L_p})\cdot(e^{\frac{qV}{k_bT}} - 1).}$$

$$\boxed{|J_{tot}| = q(n_{p0}\cdot\frac{L_n}{\tau_n} + p_{n0}\cdot\frac{L_p}{\tau_p})\cdot(e^{\frac{qV}{k_bT}} - 1)}$$

$$\implies |J_{tot}| = J_0\cdot(e^{\frac{qV}{k_bT}} - 1),\ \text{Diode equation.} \qquad (18.11)$$

Fig. 18.17 shows a plot of Equation 18.11 for two semiconductors, one of bandgap 1.1 eV (e.g. silicon), and the other 3.4 eV (e.g. GaN). The pn junction current $J_{tot} = J_0\cdot(e^{qV/k_bT} - 1)$ may look similar to the Schottky diode equation, but the contents of the coefficient J_0 are fundamentally different, leading to physically distinct behavior of the two types of diodes. The second boxed equation shows the current in a Drude-like form $J_n \sim qnv_n$, where $n = n_{p0}(e^{qV/k_bT} - 1)$ is the minority carrier density, and $v_n = L_n/\tau_n$ is an effective diffusion velocity.

This highlights the first major difference between the pn and Schottky diodes: the pn diode is a **minority carrier** device, whereas the Schottky diode is a majority carrier device. Since $n_{p0} = n_i^2/N_A \propto e^{-E_g/k_bT}$, the current density is of the form $J_{tot} \sim e^{-\frac{E_g}{k_bT}} \cdot (e^{\frac{qV}{k_bT}} - 1)$, which shows that the pn diode turns on around $V \sim E_g$, close to the bandgap of the semiconductor, as seen in Fig. 18.17. Compare this to $J_{tot} \sim e^{-\frac{q\phi_b}{k_bT}} \cdot (e^{\frac{qV}{k_bT}} - 1)$, which turns on around the barrier height ϕ_b for the Schottky diode as we have seen in Fig. 18.6. The reverse-bias current density is exponentially smaller for the wider bandgap semiconductor.

The pn junction current expression of Equation 18.11, first derived by Shockley[7], explained quite accurately the measured characteristics of high quality single-crystal germanium pn diodes produced after a prolonged effort by Gordon Teal (working with J. Little and M. Sparks) at Bell Labs. This agreement between the theory and experiment led immediately to the experimental realization of the first **bipolar junction transistor** proposed theoretically by Shockley, the subject of the next section. However, it was soon found that Equation 18.11 could not model the characteristics measured for silicon pn diodes quantitatively, even though the qualitative features were captured. The reason for this is that an additional component of current very sensitive to the bandgap has been neglected in the derivation. This component is schematically shown in the bottom of Fig. 18.15.

If there is significant generation or recombination *inside* the depletion region of the pn diode, then an additional density J_{rg} must flow for the same applied voltage. The net generation/recombination rate was derived by **S**hockley, **R**ead, and **H**all and is referred to as the **SRH** rate, given in a simplified form by[8]

$$R - G = \frac{n \cdot p - n_i^2}{\tau_0[n + p + 2n_i \cosh\left(\frac{E_t - E_i}{k_bT}\right)]}, \tag{18.12}$$

where τ_0 is a lifetime, E_t is a trap level, and E_i is the intrinsic energy level close to the middle of the bandgap. If $n \cdot p > n_i^2$, there are more carriers than thermal equilibrium, and recombination processes try to reduce their number towards $n \cdot p \to n_i^2$ by removing them from the bands. If $n \cdot p < n_i^2$, then generation processes create more inter-band carriers in an attempt to bring the system back to equilibrium. Consider the situation[9] when $n \cdot p << n_i^2$ inside the depletion region. Under this condition, $n, p << n_i$ and assuming a mid-bandgap trap level $E_t = E_i$, the generation rate and resulting current density inside the depletion region then is

$$G \sim \frac{n_i}{\tau_0} \implies J_{rg} = q \int dx(G - R) \approx \frac{qn_iW}{\tau_0}, \tag{18.13}$$

---

[7] The physicist B. Davydov in 1938 had put forth a model that was close to Shockley's 1949 model. This early model had introduced the idea of minority carrier lifetimes, but majority carrier densities were also related to the lifetime.

[8] Each term of the left hand side (LHS) and right hand side (RHS) of this equation may vary with x. The general form of this equation reads $R(x) - G(x) = ...$

[9] Since $n(x) \cdot p(x) = n_i^2 e^{(F_n - F_p)/k_bT} = n_i^2 e^{qV/k_bT}$ inside the depletion region, under a forward bias voltage $e^{qV/k_bT} >> 1$ and therefore $n \cdot p >> n_i^2$, leading to recombination. Under reverse bias condition, $qV/k_bT << -1$ and therefore, $e^{qV/k_bT} \to 0$, leading to $n \cdot p << n_i^2$, which results in generation.

where W is the thickness of the depletion region. This generation current *adds* to the total current density of Equation 18.11:

$$J_{tot} = \underbrace{qn_i^2(\frac{1}{N_D} \cdot \frac{D_n}{L_n} + \frac{1}{N_A} \cdot \frac{D_p}{L_p}) \cdot (e^{\frac{qV}{k_bT}} - 1)}_{\text{diffusion} \propto n_i^2} + \underbrace{\frac{qn_iW}{\tau_0}}_{\text{generation} \propto n_i} . \quad (18.14)$$

The diffusion current is proportional to $n_i^2 \sim e^{-E_g/k_bT}$, whereas the generation term is proportional to $n_i \sim e^{-E_g/2k_bT}$. In a narrow bandgap semiconductor pn diode, the n_i^2 term can dominate. But for a wider bandgap semiconductor pn diode, n_i^2 is rather low and the n_i term dominates, explaining the difference between Ge and Si (and other wider bandgap) pn diodes. Exercise 18.6 explores this difference further.

With modern methods of epitaxial growth of semiconductor structures, the thicknesses of the p and n layers can be controlled at will. This allows for an additional design parameter for obtaining desired diode current densities. While Equation 18.11 was derived for the case of very long undepleted p and n regions $W_p >> L_n$ and $W_n >> L_p$ (see Fig. 18.15), let us consider the general case. **The excess minority carrier concentration at a metal/semiconductor ohmic contact is zero**[10]. This property modifies the boundary conditions and the diode equation to:

$$\Delta n(x=0) = n_{p0}(e^{\frac{qV}{k_bT}} - 1) = A + B$$

$$\Delta n(x=W_p) = 0 = A \cdot e^{+\frac{W}{L_n}} + B \cdot e^{-\frac{W}{L_n}}$$

$$\implies n(x) - n_{p0} = n_{p0}(e^{\frac{qV}{k_bT}} - 1) \cdot \frac{\sinh(\frac{W_p - x}{L_n})}{\sinh(\frac{W_p}{L_n})}$$

$$\implies J_n(x) = qD_n\frac{dn(x)}{dx} = -qn_{p0}\frac{D_n}{L_n}\frac{\cosh(\frac{W_p-x}{L_n})}{\sinh(\frac{W_p}{L_n})} \cdot (e^{\frac{qV}{k_bT}} - 1)$$

$$\implies J_n^{diff}(x_p) = -qn_{p0}\frac{D_n}{L_n \cdot \tanh\frac{W_p}{L_n}}(e^{\frac{qV}{k_bT}} - 1).$$

$$\boxed{|J_{tot}| = q(n_{p0}\frac{D_n}{L_n \cdot \tanh\frac{W_p}{L_n}} + p_{n0}\frac{D_p}{L_p \cdot \tanh\frac{W_n}{L_p}}) \cdot (e^{\frac{qV}{k_bT}} - 1).} \quad (18.15)$$

This general expression of the diffusion current looks identical to Equation 18.11, except for the tanh(...) terms in the denominators. For long W_n and W_p, $W_n/L_p >> 1$ and $W_p/L_n >> 1$, and since for $x >> 1$, $\tanh x \to 1$, we recover Equation 18.11. In the opposite limit, if the undepleted regions are much smaller than the respective minority carrier diffusion lengths $W_n/L_p << 1$ and $W_p/L_n << 1$, use can use

[10] This is because at an ohmic contact, the metal can absorb or provide as many electrons or holes are needed to bring the excess density to zero.

the relation that for $x << 1$, $\tanh x \to x$, implying:

$$\boxed{|J_{tot}| = q(n_{p0}\frac{D_n}{W_p} + p_{n0}\frac{D_p}{W_n}) \cdot (e^{\frac{qV}{k_bT}} - 1),} \quad (18.16)$$

which is the short-diode equation. The current density is higher for short diodes than long diodes. In a short diode, the excess minority carrier densities change linearly (e.g.) $\Delta n(x) = n_{p0}(e^{qV/k_bT} - 1) \cdot \frac{W_p - x}{W_p}$ instead of exponentially in the undepleted regions. As a result, the minority carrier diffusion current density $J_n^{diff} = qD_n \frac{d\Delta n(x)}{dx}$ is constant in the undepleted regions, unlike the exponential decay of long diodes.

The current flowing in the pn junction in the reverse-bias condition is equally important for several applications. Fig. 18.18 shows the characteristics of a pn junction diode under reverse bias. The quasi Fermi level split is now $F_n - F_p = qV < 0$, which indicates that F_n is *below* F_p, opposite of the condition in forward bias. The minority carrier concentrations go to zero at the depletion edges because $n(x_p) \cdot p(x_p) \approx n(x_p)N_A = e^{(F_n - F_p)/k_bT} = e^{qV/k_bT} \approx 0$ because $qV/k_bT << -1$. The net current then is again given by the diffusion currents at the edges of the depletion edge:

$$J_{tot} = -q(n_{p0} \cdot \frac{D_n}{L_n} + p_{n0} \cdot \frac{D_p}{L_p}), \quad (18.17)$$

which is also obtained from the general Equation 18.11 for a negative voltage. When a minority electron in the undepleted p-side within a diffusion length L_n from the depletion edge x_p diffuses to the depletion region edge, it will experience a large electric field that pulls it into the depletion region, and ejects it out of the other side into the n-region. The energy band diagram shows this vividly: a minority electron in the conduction band on the p-side depletion edge must roll down the conduction band cliff into the n-side. The same is true for holes on the undepleted n-side. Typically the transit time of the carriers across the depletion region is much shorter than recombination lifetimes. Since increasing the reverse bias voltage simply makes the drop in E_c and E_v larger and does not increase the minority carrier concentrations at the depletion edges, the current clamps to $-J_0$ and does not depend on V, at least to first order.

However, as the reverse bias voltage is increased, the electric field maximum F_{max} at the junction as shown in Fig. 18.18 starts increasing according to

$$F_{max} = \sqrt{\frac{2qN_AN_D(V_{bi} + |V|)}{\epsilon_s(N_A + N_D)}} \underbrace{\sim}_{|V|>>V_{bi}} \sqrt{|V|}, \quad (18.18)$$

where $V_{bi} - V \to V_{bi} + |V|$ accounts for the negative sign of the voltage. When a critical electric field is reached, an electron in the conduction band can gain a kinetic energy larger than the energy bandgap of

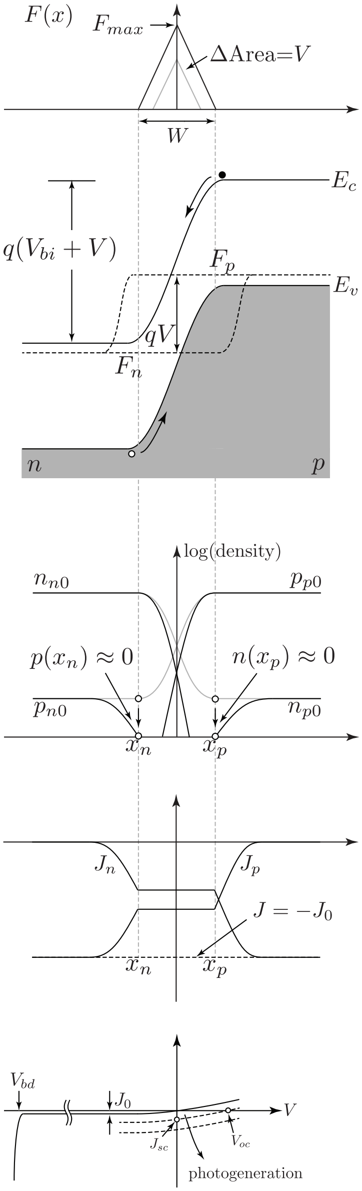


Fig. 18.18 Energy band diagram, majority and minority carrier concentrations, current components, and current-voltage characteristics of reverse-biased pn junction diodes.

the semiconductor. Then it is able to excite an electron in the valence band into the conduction band. This process starts from one electron, and produces 2 electrons and 1 hole. Each electron can now repeat this process in the conduction band, causing an **avalanche** of electrons from $1 \rightarrow 2 \rightarrow 4 \rightarrow 8 \rightarrow ....$ This sort of **avalanche breakdown** is typically observed in pn diodes when a critical electric field $F_{max} = F_{cr}$ is reached at the junction. The current increases sharply, as indicated in Fig. 18.18. The critical breakdown electric field typically depends on the doping density weakly, and very strongly on the bandgap of the semiconductor: for example, the critical field is $F_{cr} \sim 0.3$ MV/cm for silicon of $E_g = 1.1$ eV, whereas it is $F_{cr} \sim 3.3$ MV/cm for SiC and GaN of bandgaps close to $E_g \sim 3.4$ eV. If the avalanche process is allowed to proceed without restraint, removal of the electrons in the chemical bonds forming the valence band rapidly reaches such a level that the crystal structure is modified by bond breaking. This is called catastrophic breakdown of the diode, from which the diode cannot be recovered. However, the avalanche breakdown process is non-destructive if a current limit is set: this property enables several applications (see Chapter 24). For example, if a pn diode is reverse biased just short of the avalanche breakdown voltage V_{bd}, and a photon of sufficient energy to cause an interband transition is incident on it, this photon can excite an electron to trigger the avalanche process without causing destructive breakdown (see Chapter 28). The diode in this form acts as an extremely sensitive photodetector, popularly known as **avalanche photodiodes**, or APDs. APDs are able to detect single photons, and are used in secure communications and astrophysics to capture the feeble photons from distant galaxies.

The current-voltage sketch in Fig. 18.18 indicates the response of a pn diode to optical illumination. The carriers that are photogenerated inside the depletion region and in the undepleted regions move the $I - V$ curves into the fourth quadrant as indicated, leading to a short-circuit current I_{sc} and an open-circuit voltage V_{oc}. These **fourth quadrant** properties are exploited in photovoltaics, or **solar cells**, discussed further in Chapter 28. In the **third quadrant**, the reverse-bias current J_0 increases if the intensity of light is increased. If the generation rate of electron hole pairs by light is G_{op} and the recombination lifetime is τ, the minority carrier concentrations increase to $n_{p0} + G_{op}\tau$ and $p_{n0} + G_{op}\tau$. If the depletion region thickness W is much smaller than the minority carrier diffusion lengths L_n and L_p, the reverse bias current under optical illumination is increased from its dark value to

$$J_{tot} = J_{dark} + qG_{op}\tau(\frac{L_n}{\tau_n} + \frac{L_p}{\tau_p}). \tag{18.19}$$

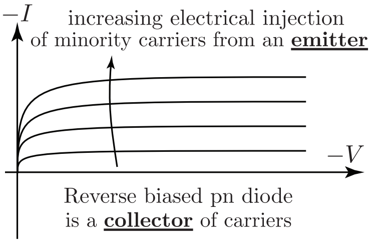


Fig. 18.19 The current-voltage characteristics of a reverse-biased pn diode under increasing electrical injection of minority carriers.

The reverse-bias current increases with increasing photogeneration, while the current is nearly independent of the voltage. This behavior will be seen if excess minority carriers are created near the depletion region by any means, not limited to optical generation.

This increase in minority carriers can be created by **electrical injec-**

tion as well. Fig. 18.19 replots the third quadrant flipped around to highlight that the reverse current increases with increasing electrical injection of minority carriers from an emitter placed nearby. A small reverse bias voltage is able to collect all available minority carriers. The reverse current does not increase beyond a certain voltage for the same reason as the dark current: the reverse bias voltage increases the potential drop and junction electric field, but does not change the minority carrier density. If one could place a minority carrier **emitter** near the reverse-biased pn diode **collector**, it is possible to achieve electrical gain. This realization led Shockley to propose, and with Sparks, Little and Teal at Bell Labs, to experimentally demonstrate the device that started the semiconductor revolution: the bipolar junction transistor.

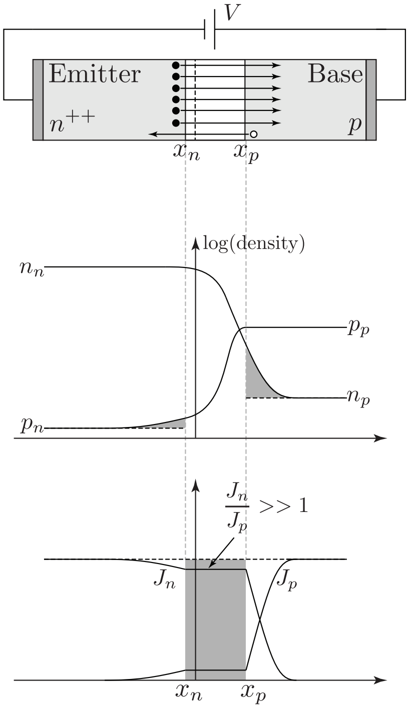


Fig. 18.20 Making an electron emitter from a pn diode. In a n^{++}/p diode, the minority electron injection is much larger than the minority hole injection. The ratio J_n/J_p is the root of electronic gain (or electronic amplification) in bipolar junction transistors.

18.4 Bipolar junction transistors

How can we construct a device that electrically injects minority carriers towards a reverse-biased pn diode? As a matter of fact we already have seen how: when a pn diode is forward biased, the current flows by the injection of minority carriers across the depletion region. However, a "symmetric" pn diode injects nearly equal number of minority carriers to both sides, leading to a nearly balanced J_n and J_p inside the depletion region and at the depletion edges, as seen in Fig. 18.15. This means only ∼half of the net current J_{tot} injects minority carriers to one side. To enhance the emitter efficiency of a single type of minority carrier to one side, the diode must be made **highly asymmetric**. Let us investigate how to design such an efficient electron emitter. Consider the ratio of the two components J_n and J_p at the depletion edge. For a long diode, this ratio is obtained from Equation 18.11:

$$\frac{J_n}{J_p} = \frac{D_n}{D_p} \cdot \frac{L_p}{L_n} \cdot \frac{n_{p0}}{p_{n0}} = \frac{L_n}{L_p} \cdot \frac{\tau_p}{\tau_n} \cdot \frac{n_{ip}^2}{n_{in}^2} \cdot \frac{N_D}{N_A}. \tag{18.20}$$

Consider the case when the diffusion lengths and minority carrier lifetimes are similar for both electrons and holes. For a homojunction pn diode the intrinsic carrier concentrations are the same on both sides of the junction $n_{ip} = n_{in}$. Then, the ratio of the two currents is simply the ratio of the doping densities $J_n/J_p \approx N_D/N_A$. Fig. 18.20 shows such a diode where the n-side is doped much heavier than the p-side by ensuring $N_D >> N_A$, indicated by the sign n^{++}. The n-side is called the **emitter**, and the p-side is called the **base**, in anticipation of their roles in the bipolar transistor. The minority electron density injected from the emitter into the base is much larger than the minority hole density injected from the base into the emitter. The corresponding ratio of current densities is given in general by Equation 18.20, and for the specific case discussed here is simply N_D/N_A. For example, if the doping densities are $N_D = 10^{17}/\text{cm}^3$ and $N_A = 2 \times 10^{15}/\text{cm}^3$, $J_n/J_p = 50$. This means the electron current at the depletion region edge is 50 times larger than the hole current. The ratio is maintained over a large range of forward bias voltages. If the two components J_n and J_p were accessible individually, then injecting say two holes from the base to the emitter will force the injection of $2 \times 50 = 100$ electrons from the emitter to the base. This is a 50X **amplification** of the signal, resulting purely from the way the diode is constructed.

However, in a two-terminal diode, the two components J_n and J_p cannot be accessed individually since the net current flowing outside the diode is always the sum of the two. Splitting the current components is achieved if a reverse biased pn-junction **collector** is merged with the n^{++}/p emitter diode, as shown in Fig. 18.21. It has now become a three-terminal device, a npn **bipolar junction transistor**. Let us discuss the operation of the device qualitatively before evaluating its quantitative characteristics. All metal contacts to the three semiconductor regions – the emitter, base, and collector are ohmic. When the emitter/base junction is forward biased with voltage $V_{BE} > 0$, a large number of electrons are injected from the n^{++} emitter into the p-type base, where they are minority carriers. The thickness of the base is intentionally designed to be far smaller than the electron diffusion length: its undepleted thickness is $W_B << L_n$. As a result, almost all the injected electrons cross the undepleted base and make it to the base depletion edge on the collector side. Very few, if any, recombine with the majority holes in the base (the figure schematically shows one out of six electrons is lost to recombination in the base).

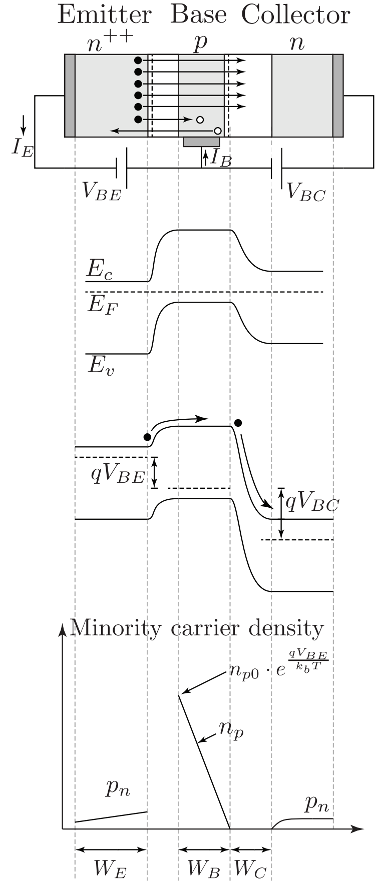


Fig. 18.21 The bipolar junction transistor. The carrier flow inside, the lower energy band diagram, and the minority carrier density profile inside the transistor is indicated for a forward-biased emitter/base junction, and a reverse-biased base-collector junction. The direction of current flow in the external wires is also indicated in response to the forward base/emitter voltage V_{BE} and a reverse base-collector voltage V_{BC}. The emitter and base thicknesses are much smaller than the respective minority carrier diffusion lengths in them.

If the base-collector pn junction is reverse biased with $V_{BC} < 0$, the minority electrons incident from the left side of its depletion region are pulled in by the electric field and ejected into the collector side. Equivalently, they roll down the conduction band edge and enter the collector, seen in Fig. 18.21. **As a result, almost all of the electron component of the current J_n ends up primarily flowing out into the collector, and not the base, even though the voltage that injects this current is applied at the base and not the collector.** The base injects a very low density of minority holes into the emitter. The net base current is dominated by this component, in addition to the very few holes for any recombination occurring in the base. Thus, the base current I_B is mostly due to J_p, and the collector current I_C is mostly due to J_n, both of the emitter-base diode. The asymmetry of the emitter/base diode therefore qualitatively leads to

$$\frac{I_C}{I_B} \approx \frac{J_n}{J_p} >> 1 \implies \text{Current gain.} \tag{18.21}$$

A current I_B flowing into the base terminal of the BJT forces an amplified current I_C that flows into the collector terminal. The current gain $I_C/I_B >> 1$ is determined by the semiconductor properties and device geometry and **not on the applied voltages or currents over a wide range**, as we will shortly discuss. Note that the current flowing in the collector terminal I_C becomes independent of V_{BC}, it depends on V_{BE} instead, again showing the **trans**ferred res**istor** effect, which is why it is called the transistor. Fig. 18.19 anticipated this property.

Bipolar transistor performance boosters: short bases and heterojunctions: In modern bipolar transistors, the thickness of the undepleted emitter W_E, and the undepleted base W_B are much smaller than the respective minority carrier diffusion lengths in those layers (see Fig. 18.21). The collector depletion region W_C is also kept short: these small dimensions make the device faster. In addition, we will now see that choosing the bandgap of the emitter E_{gE} to be larger than that of the base E_{gB} significantly boosts the transistor performance. With these modifications, the ratio of the electron current J_{nE} injected by the emitter into the base, and the hole current J_{pE} injected by the base into the emitter becomes

$$\frac{J_{nE}}{J_{pE}} = \frac{D_{nB}}{D_{pE}} \cdot \frac{W_E}{W_B} \cdot \frac{N_{DE}}{N_{AB}} \cdot \frac{n_{iB}^2}{n_{iE}^2} = \frac{D_{nB}}{D_{pE}} \cdot \frac{W_E \cdot N_{DE}}{W_B \cdot N_{AB}} \cdot e^{\frac{E_{gE}-E_{gB}}{k_b T}}, \tag{18.22}$$

where the diffusion constants and doping densities of the respective emitter/base regions are indicated by the subscripts. Note that since the bandgap difference appears in the exponential, in homojunctions this factor is unity, but in heterojunctions it is a dominant control parameter. For example, if $E_{gE} - E_{gB} = 6k_b T \sim 0.156$ eV, the J_{nE}/J_{pE} ratio is boosted by $e^6 \sim 400$ due to the exponential term. Bipolar transistors that use heterojunctions to boost their performance are sometimes called **h**eterostructure **b**ipolar **t**ransistors, or **HBT**s to distinguish them from BJTs.

The **emitter injection efficiency** γ_E of the bipolar transistor is defined as the fraction of the emitter/base diode minority carrier current injected from the emitter into the base that is capable of reaching the collector (J_{nE} in this case[11])

$$\gamma_E = \frac{J_{nE}}{J_E} = \frac{J_{nE}}{J_{nE} + J_{pE}} = \frac{1}{1 + \frac{J_{pE}}{J_{nE}}} = \frac{1}{1 + \frac{D_{pE}}{D_{nB}} \cdot \frac{W_B \cdot N_{AB}}{W_E \cdot N_{DE}} \cdot e^{-\frac{E_{gE}-E_{gB}}{k_b T}}}. \tag{18.23}$$

A very small fraction of the minority electron current J_{nE} injected into the base from the emitter may recombine in the base before making it to the collector. Let this base recombination current component be J_{rB}, implying the current that makes it to the collector is $J_C = J_{nE} - J_{rB}$. The **base transport factor** α_T of the bipolar transistor is defined as the fraction of the electron current from the emitter that makes it through the base to the collector:

$$\alpha_T = \frac{J_C}{J_{nE}} = \frac{J_{nE} - J_{rB}}{J_{nE}} = 1 - \frac{J_{rB}}{J_{nE}}. \tag{18.24}$$

With these definitions, $J_C/J_E = \gamma_E \cdot \alpha_T$. From Fig. 18.21 the currents of the three terminals follow $I_E = I_C + I_B$. Assuming a uniform cross sectional area, this translates to $J_E = J_C + J_B$. Then, the **current gain** β_F of the bipolar transistor is defined as

$$\beta_F = \frac{J_C}{J_B} = \frac{J_C}{J_E} \cdot \frac{J_E}{J_E - J_C} = \frac{\frac{J_C}{J_E}}{1 - \frac{J_C}{J_E}} = \frac{\gamma_E \cdot \alpha_T}{1 - \gamma_E \cdot \alpha_T}. \tag{18.25}$$

---

[11] γ_E is therefore a measure of the "one-sidedness" or asymmetry of electron and hole minority currents of the emitter-base diode.

Let us now evaluate the base transport factor $\alpha_T = 1 - J_{rB}/J_{nE}$. Since $J_{nE} = qn_{p0}\frac{D_{nB}}{W_B}(e^{qV_{BE}/k_bT} - 1)$ where $n_{p0} = n_{iB}^2/N_{AB}$, we must find the recombination current in the base. To do so, we use Equation 18.12

$$R - G = \frac{n \cdot p - n_i^2}{\tau_0[n + p + 2n_i \cosh\left(\frac{E_t - E_i}{k_bT}\right)]}. \tag{18.26}$$

From Fig. 18.21, we observe that inside the undepleted base region of thickness W_B, $n \cdot p >> n_i^2$, so we neglect the n_i^2 term in numerator of the recombination rate. In this region, $p >> n >> n_i$, so we neglect the n and n_i terms in the denominator. Since $\tau_0 = \tau_n$ is the minority carrier lifetime, we get

$$R(x) \approx \frac{n(x)}{\tau_n} = \frac{n_{p0}}{W_B\tau_n} \cdot (e^{\frac{qV_{BE}}{k_bT}} - 1) \cdot (1 - \frac{x}{W_B}),$$

$$\implies J_{rB} = q\int dxR(x) = \frac{qn_{p0}}{W_B\tau_n} \cdot (e^{\frac{qV_{BE}}{k_bT}} - 1)\int_0^{W_B} dx(1 - \frac{x}{W_B})$$

$$\implies \boxed{J_{rB} = \frac{Q_{nB}}{\tau_n} = \frac{qn_{p0}W_B}{2\tau_n} \cdot (e^{\frac{qV_{BE}}{k_bT}} - 1)}$$

$$\implies \alpha_T = 1 - \frac{J_{rB}}{J_{nE}} = 1 - \frac{\frac{qn_{p0}W_B}{2\tau_n} \cdot (e^{\frac{qV_{BE}}{k_bT}} - 1)}{qn_{p0}\frac{D_{nB}}{W_B} \cdot (e^{\frac{qV_{BE}}{k_bT}} - 1)}$$

$$\implies \boxed{\alpha_T = 1 - \frac{W_B^2}{2D_{nB}\tau_n} = 1 - \frac{\tau_{tr}}{\tau_n} = 1 - \frac{W_B^2}{2L_{nB}^2}}. \tag{18.27}$$

We have defined the **base transit time** for minority carriers as $\tau_{tr} = W_B^2/2D_{nB}$, and identify Q_{nB} as the total minority carrier sheet density stored in the base $Q_{nB} = \int_0^{W_B} dx \cdot \Delta n_p(x)$. Note that the undesired base recombination current J_{rB} has the same voltage dependence[12] as the desired emitter injection current J_{nE}. The ratio of the transit time to the lifetime appears explicitly. As an example, if $W_B = 100$ nm and $D_{nB} = 100$ cm^2/s, the transit time is $\tau_{tr} = W_B^2/2D_{nB} = 5 \times 10^{-13}$ s, or 0.5 ps. Compare this to typical minority carrier lifetimes τ_n, which can range from ms – ns. Since $\tau_{tr} << \tau_n$, i.e., the transit time of electrons across the base is much smaller than the time needed to recombine in the base, most of them escape into the collector. Thus, since $\tau_{tr} << \tau_n$, the base transport factor $\alpha_T \approx 1$.

Fig. 18.22 shows a npn bipolar transistor and its circuit symbol. Let us see how the device responds if the emitter is grounded, a forward bias V_{BE} is applied to the base, and a reverse bias V_{CE} is applied to the collecter. All voltages are referenced to the grounded emitter. The collector current density J_C increases as $\exp(qV_{BE}/k_bT)$, as does the base current J_B, increasing by 10 times every 60 mV at room temperature. The ratio J_C/J_B is the gain β_F. The value given by Equation 18.25 is indeed observed to be constant over several orders of magnitude of J_C, labeled as the usable operating window of the bipolar

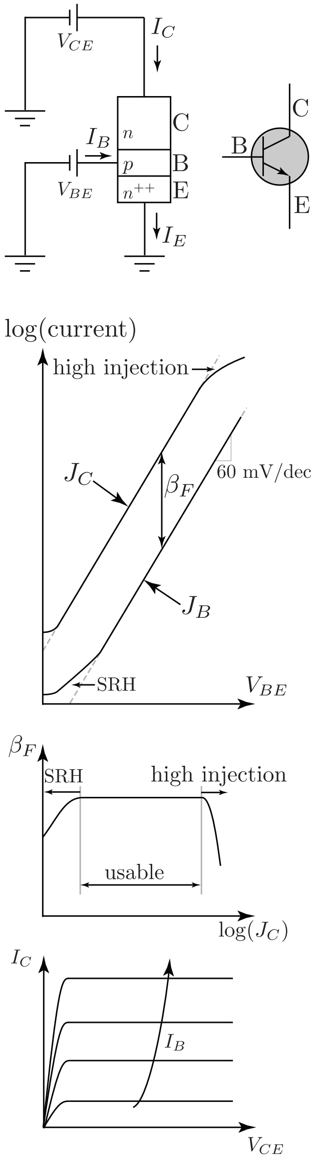


Fig. 18.22 A bipolar transistor biased in the common-emitter configuration and its symbol. The collector current J_C and base current J_B increase exponentially with V_{BE}. The ratio $\beta_F = J_C/J_B$ is the current gain. The transistor characteristics I_C as a function of V_{CE} of the common-emitter circuit show current saturation, and its control with the base current I_B.

[12] The recombination current that occurs due to SRH processes inside the emitter-base *depletion* region goes as $e^{qV/2k_bT}$. See Exercise 18.7 to explore this further.

transistor. For very low currents, Shockley-Read-Hall recombination processes change the base current behavior to $J_B \sim \exp(qV_{BE}/2k_bT)$, reducing the gain. At high V_{BE}, the minority carrier density injected into the base from the emitter approaches the base majority carrier density. This high injection effect causes the increase of J_C with V_{BE} to roll off, reducing the gain β_F. In the usable window, the collector current I_C saturates with V_{CE}, but depends linearly on I_B via $I_C = \beta_F I_B$. The current saturation is a vital feature of the transistor: it enables one transistor to be cascaded to several others to form usable circuits.

Let us consider some design rules for the most desirable current gain of the transistor $\beta_F = \gamma_E \cdot \alpha_T/(1 - \gamma_E \cdot \alpha_T)$. Clearly, the closer the product $\gamma_E \cdot \alpha_T$ is to 1, the higher the gain. Let us first consider how to make the emitter injection efficiency γ_E given by Equation 18.23 close to unity[13]. The factor in the denominator of $\gamma_E = 1/(1 + J_{pE}/J_{nE})$,

$$\frac{J_{pE}}{J_{nE}} = \frac{D_{pE}}{D_{nB}} \cdot \frac{W_B \cdot N_{AB}}{W_E \cdot N_{DE}} \cdot e^{-\frac{E_{gE}-E_{gB}}{k_bT}}, \tag{18.28}$$

should be made as small as feasible. This is achieved by reducing the base thickness W_B and base doping N_{AB} while increasing the emitter doping N_{DE}. Though W_E can be increased, it is preferable to keep it small to ensure high speed operation. However, there are design tradeoffs that one must wrestle with: if the base is made too narrow, the lateral resistance is too high, which adversely affects the gain and speed. If the emitter is doped too heavy, the bandgap shrinks due to many-particle effects. Several of these limitations are significantly countered if heterojunctions are used. If the emitter bandgap is larger than the base, the exponential factor can allow heavy base doping and lighter emitter doping, and still maintain a high γ_E. Typically in modern bipolar transistors, the base transport factor α_T is closer to unity than the emitter injection efficiency γ_E. However, if we assume $\gamma_E = 1$, then since $\alpha_T = 1 - W_B^2/(2D_{nB}\tau_n)$, the transistor gain becomes

$$\beta_F \approx \frac{\alpha_T}{1-\alpha_T} \approx \frac{\tau_n}{\frac{W_B^2}{2D_{nB}}} - 1 \approx \frac{\tau_n}{\tau_{tr}} = \frac{\text{lifetime}}{\text{transit time}}, \tag{18.29}$$

where we have neglected the factor of 1 since $\tau_n >> \tau_{tr}$. The fact that **electronic gain is the ratio of the lifetime to transit time** is a powerful general concept for minority carrier devices[14].

Speed of bipolar transistors: The transistor gain we have evaluated is its value at the low-frequency, or DC limit. If the base current is varied rapidly in time at frequency f, the collector current is amplified only upto a certain f. The speed of the transistor is quantified by the unity gain cutoff frequency f_T and the power gain cutoff frequency f_{max}. If instead of a purely DC current, an additional AC current was injected into the base $I_B + i_b \cos(2\pi ft)$ of frequency f Hz, the collector current should be $\beta_F(f)(I_B + i_b \cos(2\pi ft))$. However, $\beta_F(f)$ becomes frequency dependent, and drops with increasing f, reaching unity (= no gain) at the frequency f_T. The electronic gain decreases because it takes time for minority carriers to enter or leave the base and emitter, or for them to traverse the depleted collector region. The transistor speed is limited by a combination of all possible delays. For example, the delay incurred to charge/discharge the base is simply $\tau_B = Q_{nB}/J_{nE} = W_B^2/2D_{nB}$, or the base transit time. Similarly, the delay associated with electrons to traverse the depleted region of the collector/base junction is $\tau_C = W_C/2v_{sat}$, where $v_{sat} \sim 10^7$ cm/s is the saturation velocity of electrons in the base/collector depletion region[15]. Combining all the delays using $|\beta_F(f_T)| = 1$, one obtains

---

[13] Note that if $\alpha_T = 1$, the gain is $\beta_F = \gamma_E/(1-\gamma_E)$. Conversely, if $\gamma_E = 1$, the gain becomes $\beta_F = \alpha_T/(1-\alpha_T) = (2D_{nB}\tau_n/W_B^2) - 1$ as seen below.

[14] The meaning of this relation is the following: if we inject one hole into the base of the transistor, it must linger around for τ_n seconds before recombining with a minority electron. However, since a minority electron is whizzing past the base into the collector, it spends only τ_{tr} seconds in the base, where $\tau_{tr} << \tau_n$. For example, $\tau_{tr} \sim$ ps, but $\tau_n \sim \mu$s. If the emitter only injected one electron in response to the one hole in the base, it will only spend τ_{tr} seconds in the base, and for the rest of the time the hole will make the base not charge neutral. Therefore, to balance the hole, the emitter must inject τ_n/τ_{tr} electrons so that in the base there is one hole, and *on an average* one electron with which the hole will recombine. The concept that gain is equal to the ratio of lifetime to transit time also appears in high speed photodetectors and other photonic devices.

$$\frac{1}{2\pi f_T} = \underbrace{\frac{1}{\beta_F}\frac{W_E^2}{2D_{pE}}}_{\tau_E} + \underbrace{\frac{W_B^2}{2D_{nB}}}_{\tau_B} + \underbrace{\frac{W_C}{2v_{sat}}}_{\tau_C} + \underbrace{\frac{C_{JE}}{g_m}}_{\tau_{EB}} + \underbrace{(\frac{1}{g_m} + R_E + R_C)C_{JC},}_{\tau_{BC}} \qquad (18.30)$$

and the power gain cutoff frequency is

$$f_{max} = \sqrt{\frac{f_T}{8\pi R_B C_{JC}}}. \qquad (18.31)$$

Here, $g_m = I_C/(k_bT/q)$, $C_{JE} = C_E^{depl} + C_E^{diff}$ is the sum of the depletion and diffusion capacitances of the emitter/base junction, $C_{JC} = C_C^{depl} + C_C^{diff}$ of the collector/base junction, and R_E and R_C are geometric resistances of the emitter and the collector drift regions, inclusive of contact resistances. In modern high speed bipolar transistors, the typical delay time for each of the five terms listed is of the order of $\sim$ ps, which leads to a $f_T \sim 30$ GHz. By careful design of the semiconductor materials, doping profiles, heterostructures (e.g. in SiGe-based HBTs and in InP based HBTs), and device geometries, the speed in modern bipolar transistors have reached cutoff frequencies well in excess of 300 GHz, and some have even hit the 1 THz mark – all at room temperature!

The maximum amount of electrical power the transistor can amplify is $P_{max} = \eta I_C \cdot V_{CE}$ where $\eta < 1$, based on the way the device is biased[16]. It is desired to have the highest I_C and the highest V_{CE} to enable the transistor to amplify the most power. The amplification of power should be achieved for the highest frequencies of signal applied to the base via I_B. As may be seen from Equation 18.30, the device can be made faster by making W_B, W_E, W_C smaller and by pushing a high current through it. However, if W_C is made smaller, the reverse-biased collector/base pn junction breaks down at a smaller V_{CE} by the avalanche mechanism, limiting the total output power. This sort of tradeoff is called the **gain-bandwidth product** limitation of a transistor technology. The gain-bandwidth product limitation is related to the physical properties of the semiconductor itself [17], and can be improved by changing the semiconductor material itself.

Example 18.1 Gain and speed of a bipolar transistor

Consider a semiconductor npn BJT with $N_{DE} = 10^{17}/\text{cm}^3$, $W_E = 300$

---

[15] Intuitively, one may think the collector delay should be W_c/v_{sat}. However, since this is a depletion region, τ_C is the time it takes to extract mobile charge from the width W_C. Since for a charge at x it takes $\tau(x) = \frac{W_C - x}{v_{sat}}$ to be extracted from the depletion region, on an average it takes $\tau_C = \langle \tau(x) \rangle = \int_0^{W_C} dx \cdot \tau(x) / \int_0^{W_C} dx = W_C/2v_{sat}$ to extract all the charge. This value can be extracted more rigorously using a current-control model, where $\tau_C = J_C/\Delta Q_C$ if the extra charge in the collector depletion region is ΔQ_C.

[16] The transistor takes DC power supplied to it by the bias condition $I_C \cdot V_{CE}$ and converts part of it into AC output at the same frequency as the signal applied to the base.

[17] This sort of tradeoff applies to every transistor technology, including the field-effect transistors (FETs), the topic of next chapter.

nm, $N_{AB} = 5 \times 10^{15}/\text{cm}^3$, W_B=300 nm, and $W_C = 1000$ nm. Assume $D_{nB} = 100$ cm^2/s, $D_{pE} = 50$ cm^2/s, and minority carrier lifetimes of $\tau_{nB} = 1$ μs in the base and $\tau_{pE} = 0.5$ μs in the emitter. Find the current gain of the transistor. Find the unity current gain cutoff frequency f_T of the transistor, neglecting the collector and base junction capacitance delays, and assuming the electron saturation velocity $v_{sat} = 10^7$ cm/s. Find the current flowing in the collector when the BJT is biased in the common-emitter mode with $V_{BE} = +5$ V, and $V_{CE} = +10$ V as shown in Fig. 18.23.

Solution: Since the emitter and base widths are 300 nm each, much smaller than typical minority carrier diffusion lengths L_n, L_p of several μm (e.g. in silicon), these are short diodes. Thus, the emitter injection efficiency of the BJT is $\gamma_E = 1/(1 + \frac{50}{100} \cdot \frac{300 \cdot 5 \cdot 10^{15}}{300 \cdot 10^{17}} \cdot e^{-\frac{0}{k_b T}}) \approx 0.976$. The base-transport factor is $\alpha_T = 1 - (300 \cdot 10^{-7})^2/(2 \cdot 100 \cdot 10^{-6}) \approx 1$. For this structure, the base transport factor is nearly unity, and thus the gain is $\beta_F \approx \frac{\gamma_E}{1-\gamma_E} \approx 40$. The emitter delay is $\tau_E = W_E^2/(\beta_F D_{pE} \tau_{pE}) =$ 0.2 ps, the base delay is $\tau_B = W_E^2/(D_{nE} \tau_{nE}) = 4.5$ ps, and the collector delay $\tau_C = W_C/2v_{sat} = 5$ ps. The unity gain cutoff frequency is then given by $f_T = 1/[2\pi(\tau_E + \tau_B + \tau_C)] = 16.4$ GHz.

The built-in voltage of the emitter/base pn diode is given by $V_{bi} = (k_b T/q) \ln(N_{DE} N_{AB}/n_i^2) = 0.76$ V. When the transistor is biased as shown in Fig. 18.23 and the base/emitter diode turns on, the voltage at the base contact is V_{bi} above the emitter, which is grounded. Therefore, the base current is $I_B = (V_{BE} - V_{bi})/R_B = (5 - 0.76)V/50k\Omega = 85$ μA. Thus, the collector current is $I_C = \beta_F I_B = 3.4$ mA, which is very close to the emitter current, which is the sum $I_E = I_C + I_B \sim 3.485$ mA. As a result, the voltage that develops at the collector contact is $V_{CE} - I_C R_c = 10 - (0.0034 \text{ A}) \cdot (50\Omega) \sim +9.8$ V. Note that indeed, the base/emitter p/n diode is forward biased, and the base/collector diode is reverse biased, with $V_{CB} = 9.8 - 0.76 \sim 9.0$ V.

If an AC signal is added to the base current, it will come out amplified 40 times in the collector as long as the frequency is below $f_T \approx 16$ GHz. Such high frequency amplifiers are used widely in communication systems, such as cell phones.

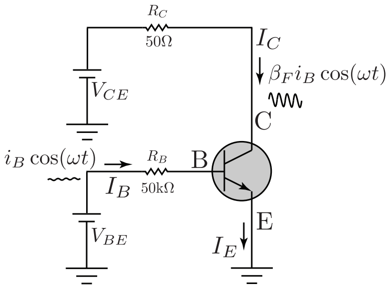


Fig. 18.23 A npn bipolar junction transistor showing the DC bias conditions, with the currents and voltage biases for Example 18.1.

18.5 Deathniums!

Prior to the first realization of the bipolar junction transistor, Bardeen and Brattain had invented the point-contact transistor in 1948, which also had an emitter, a collector, and a base (see Fig. 18.24). This transistor provided significant amplification, but there was no current saturation. The reason for the demise of the point-contact transistor was its very low reliability, primarily due to the fact that the current flow and device operation was heavily affected by defects and traps near the surface that are non-radiative recombination sites. Shockley named these deep trap states **deathniums**.

Fig. 18.24 John Bardeen (left), William Shockley (middle), and Walter Brattain (right) of the Bell Laboratories were jointly awarded the Nobel Prize in Physics in 1956 for the invention of the transistor. It is said that Brattain and Bardeen did not particularly like this photograph because it seemed to imply Shockley was the experimentalist among them, which was not the case at all!

The bipolar junction transistor started replacing vacuum tube based amplifiers in the 1950's due to its far smaller size and weight, and most importantly, its far superior reliability. The initial bipolar transistors could only amplify 10s of MHz signals and were not competitive with vacuum tubes in speed (which could amplify few GHz signals). With rapid miniaturization and the development of several new processing techniques such as the mesa and planar processes, the speed of bipolar transistors crept up and soon exceeded those of tubes, launching the field of solid state devices.

Fig. 18.25 Gordon Teal in early 1950s succeeded in growing single-crystal germanium npn junctions, enabling the first bipolar junction transistor. High quality single-crystal growth and control of Shockley's "deathniums" was critical to the operation of the device. Teal moved to Dallas and was able to achieve the same with single-crystal growth of silicon npn junctions. The silicon npn bipolar transistor launched in 1954 was the first commercially available transistor. It was a superior alternative to the Ge transistors because of higher temperature operation. It launched the ascent of the then unknown company: Texas Instruments.

Semiconductor researchers till today have always had to wrestle with controlling Deathniums and keeping their number low enough to make operational devices. Though the name deathnium has been lost over time, it aptly describes the role of both surface and bulk traps, so we will use the name. The role of deathniums is to trap electrons and holes in deep energy levels in the bandgap, preventing their desired motion in the bands, reducing the desired currents, and messing up the device electrostatics by creating undesired electric fields. A simple model for deathnium-induced surface-recombination currents is $J_s = q\int_0^L dx R_s(x)$ where $R_s(x)$ is the surface recombination rate, and L is a diffusion length from the surface. If these states are not reduced, they cause surface leakage currents in the pn junction that is similar in principle to the undesired low-bias SRH bulk recombination, degrading the performance.

Noyce
Kilby

Fig. 18.26 Robert Noyce (left) and Jack Kilby (right) are the inventors of the integrated circuit (IC). The IC puts several transistors, resistors, and other electronic passive and active elements on a single piece of a semiconductor wafer to perform several functions. Noyce founded Fairchild Semiconductors, and later Intel. Kilby, of Texas Instruments, was awarded the 2000 Nobel Prize in Physics to recognize the revolutionary effect of the IC on human civilization.

The success of the bipolar transistor owes a great deal to an unsung hero: Gordon Teal (Fig. 18.25), who developed the crystal growth process that enabled the first single-crystal npn germanium bipolar transistors. However, germanium transistors were prone to a large variation in performance with temperature because of its small bandgap. The move from germanium to silicon bipolar transistors was swift: because of the larger bandgap, the silicon bipolar junction transistor offered superior performance with temperature variation. It was the first solid state device to be commercialized by Texas Instruments (Fig. 18.25). To avoid the deleterious effects of deathniums at the surface, it was soon discovered that covering the exposed surface of silicon bipolar transistors with a thermally grown silicon dioxide (SiO_2) improved the uniformity, performance, and reliability of the bipolar transistors significantly. This discovery allowed for nearly identical devices to be fabricated on a planar surface, and led researchers at Fairchild semiconductors and Texas Instruments to connect several transistors on the same planar process, leading to the first **integrated circuit** (see Fig. 18.26). It also paved the way for the device that would soon replace Si bipolar transistors: the metal-oxide field-effect transistor (**MOSFET**), which is the subject of the next chapter.

18.6 Chapter summary section

In this chapter, we learned:

- The Schottky diode allows the flow of *majority* carriers of the semiconductor across the metal/semiconductor junction, and results in a current density of the form $J = A^{\star} e^{-\frac{q\phi_b}{k_b T}} \cdot (e^{\frac{qV}{k_b T}} - 1)$. Here $A^{\star}$ is an effective Richardson coefficient that depends on the effective band mass of the semiconductor, ϕ_b is the surface potential barrier height, and V is the applied voltage across the junction. The rectification can easily exceed $\sim 10^{10}$!
- A Schottky diode can be used to rectify the direction of current flow, and also to block large reverse voltages. When reverse bias voltage is applied to the junction, the depletion region extends into the semiconductor. This property is used in field-effect transistors. The rectifying behavior of the Schottky diode can be converted to ohmic behavior if the semiconductor near the junction is heavily doped by enhancing tunneling currents over thermionic emission. This enables efficient ohmic contacts. The combination of the reverse-bias depletion, and tunneling ohmic contacts is often used to realize field-effect transistors.
- Unlike the Schottky diode, the pn junction diode is a *minority* carrier device. The current density of a pn junction diode is $J = q(n_{p0}\frac{D_n}{L_n} + p_{n0}\frac{D_p}{L_p}) \cdot (e^{\frac{qV}{k_b T}} - 1)$ for long diodes, and $J = q(n_{p0}\frac{D_n}{W_n} + p_{n0}\frac{D_p}{W_p}) \cdot (e^{\frac{qV}{k_b T}} - 1)$ for short diodes. Here D_n, D_p are the diffusion constants of minority electrons and holes, n_{p0}, p_{n0} are the minority electron and hole concentrations respectively, and W_n, W_p are the undepleted lengths of the n and p sides, which for the short diodes are much smaller than the minority carrier diffusion lengths L_n, L_p.
- The electron and hole components of the pn junction diode current can be made heavily asymmetric by doping and bandgap control. When a reverse-biased collector/base pn junction is brought very close to a forward biased n^{++}/p emitter/base junction, the electron component of the current of the forward biased diode can be collected by the collector terminal, whereas the much smaller hole component flows in the base terminal. The ratio of the collector to the base current $\beta_F = I_C/I_B$ of this 3-terminal npn dipolar transistor offers *electronic gain*.
- The performance of a bipolar junction transistor can be boosted by the use of narrow base and emitter, and controlling the doping densities so as to achieve highly asymmetric current flow of J_n and J_p. Use of heterostructures in bipolar transistors is particularly potent in allowing simultaneous high gain and high speed operation of HBTs.
- Bipolar transistors launched the field of semiconductor devices, by the combined development of the underlying physics, perfection of materials science of crystal growth, doping, and heterostructure control, and techniques of fabrication such as the planar process and the integrated circuit.

Further reading

The physics of Schottky diodes, pn junction diodes, and bipolar transistors is described well in **Device Electronics for Integrated Electronic Circuits** by Muller, Kamins, and Chan, and in **Fundamentals of Modern VLSI Devices** by Taur and Ning. Some popular undergraduate textbooks are **Semiconductor Device Fundamentals** by Pierret, and **Solid State Electronic Devices** by Streetman and Banerjee. For an historical account of the discovery and development of these devices, I highly recommend **Semiconductor Devices: Pioneering Papers**, a collection by Simon Sze. For the commercial angle, **The History of the World Semiconductor Industry** by Morris tracks the early development of the field, till the late 1980s.

Exercises

(18.1) **Appreciating the rectifying power of diodes**

The current-voltage characteristics of ideal Schottky and pn junction diodes follow $I = I_0(e^{qV/k_bT} - 1)$. For Schottky diodes, the coefficient $I_0 = I_{Schottky}e^{-q\phi_b/k_bT}$ where $q\phi_b$ is the Schottky barrier height. For pn junction diodes, $I_0 = C_{pn}e^{-qV_{bi}/k_bT}$ where $V_{bi} = \frac{k_bT}{q}\ln(\frac{N_aN_d}{n_i^2})$ is the built-in voltage.

(a) Verify these dependences from the discussion in the chapter.

Answer the following questions to appreciate the rectifying power of both Schottky and pn junction diodes, and to also realize the differences between the two type of diodes.

(b) Make a plot of $I/C_{Schottky}$ as a function of voltage V for $T = 300$ K with the y-axis and x-axis in the linear scale for $q\phi_b = 1.0$ eV . What is the turn-on voltage? Note that the turn-on voltage may appear different when plotted over different ranges of forward voltages. However, the maximum experimental currents are typically limited not by the diode but by voltage drops in external resistances.

(c) Replot the absolute value of the current $I/C_{Schottky}$ vs. the voltage V with the y-axis in log-scale and V in linear scale. What is the rectification ratio $\frac{I_F}{I_R} = \frac{I(V{=}{+}1.0\text{Volt})}{I(V{=}{-}1.0\text{Volt})}$? Note that this can exceed 10^{15} easily: this is nearly a million times the population of the earth! In other words, for every electrons allowed to flow at that reverse bias, nearly $\geq 10^{15}$ electrons flow in the forward bias direction!

(d) Examine the primary reason for this one-way traffic control. Argue that all bands and electron states in normal semiconductors allow electron flow both ways equally well, but the source of electrons going one way versus the other is the origin of the rectification.

(e) Compare and contrast the rectification of a Schottky diode to that of a pn junction diode. Note that typically the Schottky barrier height is only a small portion of the bandgap of the semiconductor. This should make the pn junction current density far smaller than a Schottky diode made of the same semiconductor. Verify this is the case.

(f) Compare and contrast the speed at which a Schottky diode can be made to switch compared to a pn junction diode. The minority carriers of pn junction diodes cause delays that are absent in Schottky diodes.

(g) Compare and contrast the breakdown voltages in reverse-bias that can be sustained in a Schottky vs. a pn junction diode. The breakdown mechanisms are discussed in Chapter 24.

(h) Typically a pn junction diode sustains higher breakdown voltages, but lower on-currents than a Schottky diode based on the same semiconductor. Argue why this must be the case based on what you have learned in this problem.

(18.2) The force is not as strong: image force lowering in Schottky diodes

(a) The lowering of the Schottky barrier height by the image force depends strongly on the dielectric constant ϵ_s of the semiconductor. Explain why it is difficult to form a Schottky barrier with a large barrier height and large rectification for semiconductors with low dielectric constants.

(b) The image force lowering of the Schottky barrier is observable only in processes that have a current flowing. For example, the barrier lowering is not observable in a quasi-static capacitance-voltage measurement. Argue why this must be the case because only an electron that is in transit through the depletion region of the diode feels the image force attraction to the metal. An electron inside the metal, or inside the conductive region of the semiconductor does not feel this force.

(18.3) Squeeze and stretch: Schottky currents in semiconductors with multiple anisotropic valleys in the bandstructure

The Schottky diode current was evaluated by summing over the states in the conduction band at the top of the barrier in the semiconductor. The conduction band edge of a direct-bandgap semiconductor such as GaAs or GaN has an isotropic effective mass with a single valley $g_v = 1$ at the Γ-point. But for semiconductors such as Si or Ge, there are multiple valleys that are ellipsoidal. Discuss why the sum over valleys should depend on the crystal direction of the metal/semiconductor plane, and how the ellipsoidal geometry affects the net current in the diode. If strain is applied to modify the shape of the ellipsoidal bands, how will the Schottky diode current change?

(18.4) Not at the top: evaluating ballistic currents inside the semiconductor

Let us evaluate the current in a Schottky diode not at the top of the barrier, but at the depletion edge plane $z = z_d$ deep inside the semiconductor in Fig. 18.6 (a). In this plane, there are electrons moving to the left towards the metal, and those moving to the right towards the bulk of the semiconductor. Among those moving to the left towards the metal must be incident from the bulk of the semiconductor: since there is no scattering in the ballistic approximation, this is the only possibility. But the right going electrons in this plane are composed of two types: those incident from the metal that made it *over* the barrier, and those that were initially moving to the left incident from the bulk, but whose energies were lower than the barrier height $q\phi_b$, that were quantum mechanically *reflected* from the barrier and are now returning to the bulk. Note that this is still ballistic. Since the right going electrons are partly in equilibrium with the metal, and partly with the semiconductor, calculating the current in this plane is a bit more involved compared the case of a plane where both populations are in equilibrium with only one electrode each. Furthermore, there are some electrons that can tunnel through the barrier, which further causes a mixing of the two electrodes. Pursue this strategy to examine how current continuity must be maintained across the entire diode.

(18.5) Quantum current across a Schottky junction

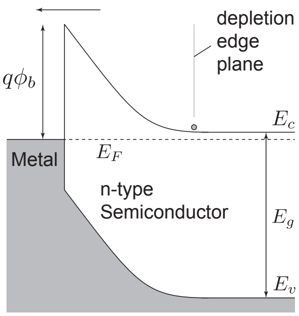


Fig. 18.27 Schottky diode.

The energy band diagram in Fig. 18.27 is of a metal/semiconductor junction of surface barrier height $a\phi_b$.

(a) If the work function of the metal is $q\phi_M$, what is the work-function of the semiconductor? Why is there a depletion region in the semiconductor?

(b) Sketch the distribution of electrons in the conduction band at the depletion edge plane indicated as a function of energy.

(c) The semiconductor conduction bandstructure is given by $E(k_x, k_x, k_z) = E_c + \frac{\hbar^2}{2m_c^\star}(k_x^2 + k_y^2 + k_z^2)$ where E_c is the band-edge energy, and $m_c^\star$ is the conduction band effective mass.

(d) The expression for the Schottky diode current is $J = J_0(e^{\frac{qV}{k_bT}} - 1)$. Make a linear and a log-scale plot of J/T^2 vs. voltage for a GaAs-metal Schottky diode of barrier height $q\phi_b = 1$ eV and $m_c^\star \sim 0.067m_0$ at T =300K. In the log-plot, indicate the y-axis intercept. Argue why you can use such a plot to experimentally measure the conduction band effective mass $m_c^\star$. How would you measure ϕ_b?

This problem has been solved over the last century by Walter Schottky, Owen Richardson, Arnold Sommerfeld, Ralph Fowler, Lothar Nordheim, Hans Bethe, and Nevill Mott, among others. The Schottky diode rectifier is an extremely important semiconductor device, and the rectifying property was one of the earliest to be used in crystal radios, and the earliest pointers to the existence of semiconductors – you can imagine why people have been interested in it for so long!

(18.6) Non-idealities in pn diodes

Consider a pn junction diode with doping densities $N_D = N_A = N_0$, minority carrier diffusion lengths $L_n = L_p = L_0$, and lifetimes $\tau_n = \tau_p = \tau_0$ under reverse bias. Show that the ratio of generation/recombination to diffusion currents is

$$\frac{J_{rg}}{J_{diff}} \approx \frac{W}{L} \cdot \frac{N_0}{n_i}. \qquad (18.32)$$

Explain under what conditions the recombination/generation current is dominant over the diffusion current. Argue why for narrow bandgap semiconductors, the diffusion current J_{diff} can be much larger than J_{rg}, making the ideal diode equation accurate, whereas for wider bandgap semiconductors should show increasing deviation from the ideal diode model.

(18.7) Half-Boltzmann currents at low bias

At low forward bias, Shockley-Read-Hall (SRH) recombination of electrons and holes inside the depletion region of a pn junction diode causes the diode current to depend on the voltage as $I \sim e^{qV_{BE}/2k_bT}$, where there is a factor of 2 in the denominator of the exponent. To see why this is the case, start from the general SRH recombination rate expression of Equation 18.12, and use the splitting of the quasi Fermi level in forward bias $F_n - F_p = qV_{BE}$ to obtain an expression for the recombination current. Then assuming the trap level $E_t \approx E_i$, find the extrema of the denominator to find the maximum recombination rate. You should obtain a $e^{qV_{BE}/2k_bT}$ in the denominator of the recombination rate, which combined with the numerator gives the half-Boltzmann slope of the diode turn-on at very low currents.

Zeroes and Ones: The Ballistic Transistor

19

19.1 The MOS capacitor 430
19.2 The ballistic FET 433
19.3 Ballistic I-V characteristics 436
19.4 Quantum wire ballistic FET 441
19.5 The drift-diffusion FET 443
19.6 CMOS and HEMTs 445
19.7 Source/drain ohmic contacts 451
19.8 A brief history of FETs 452
19.9 Chapter summary section 454
Further reading 455
Exercises 455

In the last chapter, we learned the principle of operation of bipolar transistors. In this chapter, we apply the quantum mechanics of electrostatics and charge transport in semiconductors to understand the characteristics of *unipolar* field-effect transistors (FETs). We first consider the situation when transport of electrons in the FET occurs without scattering, i.e., ballistically from the source contact to the drain contact. The study of the ballistic FET will highlight various quantum limits of performance of the device, guiding the optimal choices of semiconductor and dielectric materials, and the device geometry to extract the highest performance from such devices. We then explore how the FET behavior changes with scattering, using a drift-diffusion model of charge transport. Two of the most commonly used FETs, the CMOS and HEMT technologies are discussed. We end the chapter with a historical account of the development of FETs. Through this chapter, we develop powerful insights into the inner workings of the remarkable device that powers the digital world by learning:

- How the field-effect mechanism dynamically controls mobile electron and hole densities in semiconductors,
- The quantum mechanical operation principles of FETs,
- Fundamental quantum limits of operation of ballistic FETs,
- Effect of scattering and drift/diffusion transport on the operation of FETs, and
- How FETs in today's CMOS and HEMT technologies are used for logic, memory, and communications, and the historical development that led to the current state of the art FET technologies.

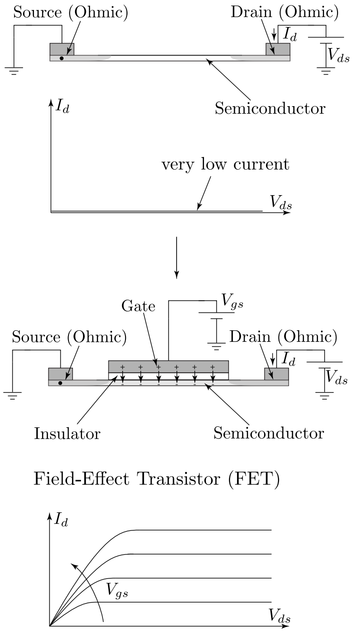


Fig. 19.1 A high-resistance semiconductor is converted to a field-effect transistor (FET) by inducing mobile carriers by field effect caused by a gate voltage.

The core working principle of a FET is simple to understand, as illustrated in Fig. 19.1, which represents just one of many embodiments of the device. Consider a thin semiconductor to which ohmic contacts are made to *one* of the bands: say the conduction band. The source and drain regions under the metal have been doped heavily with donors to make them n^{++} for low-resistance tunneling contacts. The rest of the semiconductor is left undoped, which means $E_F << E_c$, and there are very few electrons in the conduction band in this region, consistent with $n \sim N_c \cdot \exp[-(E_c - E_F)/k_b T]$. Since the resistance of the semiconductor channel is high, a very small current $I_d \sim qnv$ where v is an effective velocity will flow if a voltage V_{ds} is applied on the drain contact. A gate metal is then brought into close proximity to the semiconductor, separated by an insulator. The insulator prevents electron flow from the semiconductor to the metal, but allows electric field lines to go through. The gate metal and the semiconductor form a parallel plate capacitor: applying a +ve voltage on the gate V_{gs} will induce negatively charged electrons in the semiconductor by field effect, flooding the conduction band with mobile carriers and reducing the resistance. We will see shortly that the mobile carrier density n in the semiconductor channel is controlled **exponentially** by the voltage on the gate metal via $n \sim N_c \cdot \exp\left[qV_{gs}/k_bT\right]$ over a range of voltages below a threshold, and then polynomially above this threshold. This will allow a substantial current $I_d \sim \exp\left[qV_{gs}/k_bT\right]$ to flow in the drain. As indicated in Fig. 19.1, above a certain V_{ds}, **the drain current I_d will depend only on the gate voltage V_{gs} and *not* the drain voltage**. A small change in the gate voltage will cause a large change in the drain current I_d, providing voltage gain, making the FET an **amplifier**. Since the state of the semiconductor channel is varied by the gate voltage from very resistive (insulating) to very conductive (metallic), the FET is also an electronic **switch**, conceptually identical to a light switch we "gate" mechanically by hand in our rooms. The FET switch is much faster than our hand, and needs far less energy because electrons are light, and need tiny voltages to move! Thus, FETs are used both to amplify signals for communications, and as digital switches for logic.

In Chapter 18, we learned how the p-type base of a bipolar transistor controlled the flow of **minority** electrons that traveled from the emitter to the collector (see Fig. 19.2). The bipolar transistor is therefore a **current-controlled** device, in which one can obtain **current gain**: the collector current is an amplified version of the base current $I_C = \beta_F I_B$. There are some similarities, and some important differences between a BJT and a FET. The first and foremost difference is: a FET is a **unipolar** device. The current is carried by majority carriers: either electrons in n-channel FETs, or holes in p-channel FETs, recombination is avoided. The second difference is: the FET is a **voltage controlled** device, which enables **voltage gain**. In a BJT, the current that flows from the emitter to the collector is controlled by the current that flows into the base. In a FET, the current that flows from the **source** to the **drain** is controlled by the voltage applied on the **gate**. We first discuss how the gate exerts control over the electrical conductivity of the semiconductor channel, assuming very low resistance source/drain ohmic contacts[1].

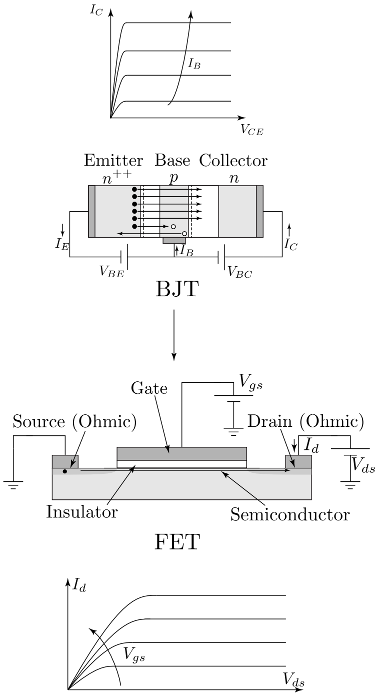


Fig. 19.2 The FET is a voltage-controlled majority-carrier device, compared to a BJT which is a current-controlled minority-carrier device. The roles of the terminals Source↔Emitter, Drain↔Collector, and Gate↔Base are analogous. In an ideal FET no current flows into the gate: the drain current is controlled through the gate voltage. In an ideal BJT, the collector current is controlled by the base current. The FET in this figure is similar to Fig. 19.1, but with a substrate that is either insulating, or has carriers in a different band than the source and drain contacts.

19.1 The MOS capacitor

Let us first investigate the quantum mechanics of the metal-oxide-semiconductor (MOS) capacitor[2]. As we have discussed qualitatively in Chapter 18 (Fig. 18.11), the gate terminal of FETs can be made of

[1] Contacts are discussed in Section 19.7.

[2] The gate insulator need not be an oxide. In III-V HEMTs, it is an epitaxially grown wider bandgap crystalline layer such as AlN in GaN HEMTs, or AlGaAs in GaAs HEMTs. We will continue calling this layer an "oxide", and the gate stack a MOS in a slight abuse of terminology, but the context will make it clear.

Schottky diodes (such FETs are called MESFETs), p-n junction diodes (called junction-FETs of JFETs), or by metal-insulator-semiconductor (MIS) structures called MISFETs, as was discussed in Chapter 16, Fig. 16.4. The insulator in traditional silicon FETs is an oxide: either SiO_2, or higher-κ dielectrics such as HfO_2 in modern integrated circuits. Thus, the gate is typically a metal-oxide-semiconductor (MOS) stack. In Chapter 17, Section 17.3, we discussed the quantum physics of the two-terminal MOS capacitor in detail: it is **strongly recommended to re-read that section** to understand the field-effect physics. For a planar capacitor, a threshold voltage was identified as $V_T = \phi_b - (\Delta E_c/q) + (E_0/q)$, where ϕ_b is the metal-insulator barrier height potential, ΔE_c is the conduction band energy offset between the insulator and the semiconductor, and E_0 is the lowest allowed eigenvalue in the conducting band of the semiconductor in Fig. 17.11.

Let us connect Fig. 17.11 to the part under the gate of the FET shown schematically in Fig. 19.3. A crucial difference between the two-terminal MOS capacitor discussed in Chapter 17 and the gate capacitor in the FET is the presence of *two* ohmic contacts instead of one: the source and the drain. These two ohmic contacts can be held at separate potentials. Let us first consider the case when both the drain and source contacts are connected to ground, and the ohmic contacts are made to the conduction band of the semiconductor, which will result in a n-channel FET. Similar arguments hold if the ohmic contacts are made to the valence band, which will result in a p-channel FET.

The voltage applied on the gate metal raises the potential of the gate metal with respect to the source by V_{gs}, and the drain by V_{gd}. When the source and drain are grounded, $V_{gs} = V_{gd}$. The semiconductor layer is then at the same potential (= ground) from the source to the drain. When a positive voltage is applied on the gate, the gate metal plate becomes positively charged. This is achieved by the battery, which extracts electrons from the gate metal and injects them into the semiconductor conduction band through the external circuit. The positive charge on the metal resides on ions that have lost their electrons within roughly a lattice constant in the metal side, at the M-O interface. Since there are two ohmic contacts in the FET, in the case of ballistic transport, the left (source) electrode injects electrons extracted from the gate metal into the semiconductor from the left side. These electron waves move from the source towards the drain, and occupy the $k_x > 0$ states in the (k_x, k_y) modes allowed in the semiconductor, which are filled up to the source Fermi level E_{Fs} in energy[3], as shown in Fig. 19.3 as the right-half circle. The right (drain) contact similarly injects electrons into the left-going modes, filling them up to its own Fermi level E_{Fd}, indicated by the left-half circle. The mode spacing for electrons in the channel are $\Delta k_y = 2\pi/W$ along the width of the FET, and $\Delta k_x = 2\pi/L$, where L is an effective S/D channel length[4]. For both ballistic and long-channel transistors, the net 2D current scales linearly with the width W for large $W >> L$.

The all-important relation between the 2D electron sheet density n_s

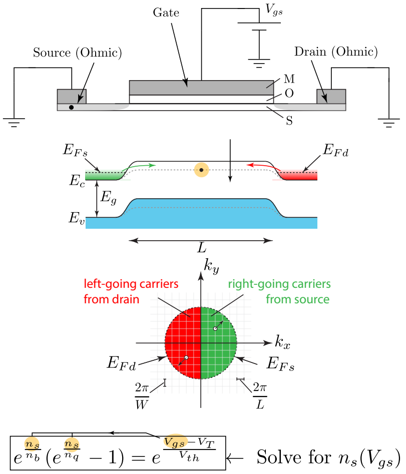


Fig. 19.3 Energy band diagram from the source to drain under the MOS capacitor. The source contact injects and fills the right-going (k_x, k_y) states with $k_x > 0$, and the drain contact fills the left-going states with $k_x < 0$ and brings the respective distributions to equilibrium with themselves by sharing the quasi-Fermi levels E_{Fs} and E_{Fd} respectively. In this situation $E_{Fs} = E_{Fd} = 0$. The transcendental equation relates the 2D mobile carrier density n_s with the gate voltage V_{gs} of the MOS capacitor.

[3] The allowed electron-wave states in the semiconductor channel can occupy the modes $(k_x, k_y) = (\pm n_x \frac{2\pi}{L}, \pm n_y \frac{2\pi}{W})$, just as the allowed modes of light in a microwave, or optical waveguide. Here $n_x, n_y = 0, \pm 1, \pm 2, \ldots$. Each state has a characteristic group velocity $v_g = \frac{1}{\hbar}\nabla_k E(k_x, k_y)$, and its occupation probability $f(k_x, k_y)$ is controlled by the contact(s) with which the state is in equilibrium.

[4] The value of the channel length L will not appear in ballistic transport, since there is no scattering in the transport from the source to the drain. However, L is rather important in scattering-dominated, long-channel transistors, where it will appear with the electron mobility. Transverse quantization and quantum-wire behavior occurs for very small W as discussed later in this chapter.

and the gate voltage V_{gs} is shown boxed in Fig. 19.3. For a 2D semiconductor channel with a parabolic bandstructure, n_s is given by

$$n_s = g_s \cdot g_v \cdot \frac{\sum_k f(k)}{W \cdot L} = \frac{g_s g_v}{WL} \int \frac{dk_y}{\frac{2\pi}{W}} \cdot \frac{dk_x}{\frac{2\pi}{L}} \cdot f(k_x, k_y)$$

$$= \frac{g_s g_v}{(2\pi)^2} \int_{-\infty}^{+\infty} dk_y \int_0^{+\infty} dk_x \cdot \frac{1}{1 + \exp\left[\frac{E_c + \frac{\hbar^2 (k_x^2 + k_y^2)}{2m_c^\star} - E_{Fs}}{k_b T}\right]} +$$

$$\frac{g_s g_v}{(2\pi)^2} \int_{-\infty}^{+\infty} dk_y \int_{-\infty}^{0} dk_x \cdot \frac{1}{1 + \exp\left[\frac{E_c + \frac{\hbar^2 (k_x^2 + k_y^2)}{2m_c^\star} - E_{Fd}}{k_b T}\right]}$$

$$\Longrightarrow \boxed{n_s = \frac{1}{2} \cdot n_q \cdot [F_0(\eta_s) + F_0(\eta_d)],} \qquad (19.1)$$

where $F_0(\eta) = \ln[1 + e^\eta]$ is the Fermi–Dirac integral of order zero, $n_q = N_c^{2d}$ is the band-edge DOS of the semiconductor channel[5], and $\eta_s = \frac{E_{Fs} - E_c}{k_b T}$ and $\eta_d = \frac{E_{Fd} - E_c}{k_b T}$. For the case $V_{gs} = V_{gd}$, $E_{Fs} = E_{Fd}$, using Equation 19.1 with the energy band diagram in Fig. 17.11, we obtain the central relation between the voltage V_{gs} on the gate metal, and the resulting mobile 2D carrier density n_s induced by field effect in the semiconductor channel:

$$\boxed{e^{\frac{n_s}{n_b}} \cdot (e^{\frac{n_s}{n_q}} - 1) = e^{\frac{V_{gs} - V_T}{V_{th}}},} \qquad (19.2)$$

where $V_{th} = k_b T / q$ is the thermal voltage, the gate insulator barrier capacitance per unit area is $C_b = \epsilon_b / t_b$, and $n_b = C_b V_{th} / q$ is an effective 2D carrier density. From Section 17.3: the dependences of n_s on V_{gs} in two limits of the gate voltage are:

$$n_s \approx \frac{C_b C_q}{C_b + C_q} \left(\frac{V_{gs} - V_T}{q}\right) \text{ for } V_{gs} > V_T \text{, and}$$

$$n_s \approx n_q e^{\frac{V_{gs} - V_T}{V_{th}}} \text{ for } V_{gs} < V_T. \qquad (19.3)$$

For the sub-threshold condition $V_{gs} < V_T$, the electron density in the semiconductor channel n_s is exponentially dependent on V_{gs}, and the quasi-Fermi level of the semiconductor $E_{Fs} = E_{Fd} < E_c$ is inside the bandgap of the semiconductor, populating it with very few electrons. On the other hand, if $V_{gs} > V_T$, the Fermi level rises to $E_{Fs} = E_{Fd} > E_c$ and the mobile carrier density is roughly linearly proportional to the excess gate voltage above threshold.

Let us compare an example of the exact solution of Equation 19.2 with the limiting cases of substhreshold and above threshold conditions of Equation 19.3. Fig. 19.4 shows, in the logarithmic scale (top), and in the linear scale (bottom), how the overdrive gate voltage $V_{gs} - V_T$ controls the mobile 2D electron gas density n_s in the semiconductor conduction band for a thin insulating dielectric. The solid

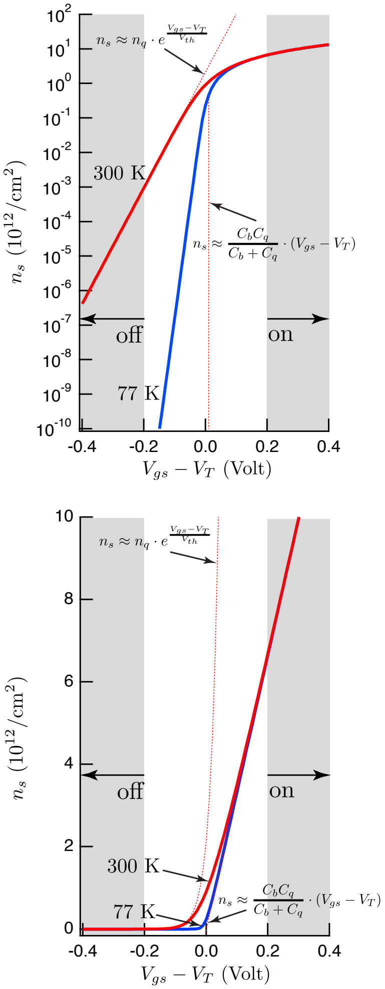


Fig. 19.4 The FET channel 2D electron gas density n_s as a function of the gate voltage $V_{gs} - V_T$, calculated at 300 K and 77 K as solutions of Equation 19.2 (solid lines), and Equations 19.3 (dashed lines). The band parameters used are an effective mass of $m_c^\star = 0.2 m_e$, spin and valley degeneracies $g_s = 2$ and $g_v = 1$. The insulator layer has a dielectric constant $\epsilon_b = 20\epsilon_0$, and a thickness $t_b = 2$ nm.

---

[5] Note that we are re-labeling $N_c^{2d} = n_q$ for this chapter, and calling it a "quantum" sheet carrier concentration to simplify the notation for 2D channel FETs. The "semi-circle" integrals of Equation 19.1 are simplified by converting to radial coordinates $(k_x, k_y) = k(\cos\theta, \sin\theta)$.

lines are the exact solutions of Equation 19.2, and the dashed lines are the subthreshold and on-state approximations of Equation 19.3. The mobile carrier density n_s is changed at 300 K from the off state value of 5×10^5 /cm^2 to $\sim 10^{13}$ /cm^2, nearly eight orders of magnitude when the gate voltage is swept over $V_{gs} - V_T = -0.4$ Volt to $+0.4$ Volt, a small interval. This will be seen later to correspond to a similar order of magnitude change in the channel conductance. **The typical on-state channel carrier concentration for most 2D FETs is $\sim 10^{13}$ /cm^2.**

It is clear that the approximations of Equation 19.3 *separately* reproduce the on-state of the channel when $V_{gs} >> V_T$ and it is very conductive, as well as the sub-threshold, or off-state of the channel when $V_{gs} << V_T$ and it is very resistive[6]. It is instructive to compare the differences at room temperature and 77 K. For example the density at $V_{gs} - V_T = -0.2$ Volt is different at the two temperatures by more than a factor of 10^7, though both densities are so low that they are negligible in the linear scale compared to the on-state densities. In the on-state, at $V_{gs} - V_T = +0.2$ Volt, the densities at 300 K and 77 K are nearly identical, and both given by roughly $n_s \approx \frac{C_b C_q}{C_b + C_q} (\frac{V_{gs} - V_T}{q})$ to an excellent approximation. Only for a small window of energies around $V_{gs} = V_T$ are the limiting approximations of Equation 19.3 not valid. We are now prepared to quantitatively understand ballistic FETs[7].

19.2 The ballistic FET

With this understanding of the gate capacitor, we proceed to discuss the operation of the ballistic FET when a drain voltage V_{ds} is applied to the drain ohmic contact. Fig. 19.5 (a) illustrates a field-effect transistor with a 2D conducting channel. The device is shown with two gates sandwiching a 2D crystal semiconductor channel, with a corresponding energy band diagram in the vertical direction shown in Fig. 19.5 (b). The two gates may have separate bias voltages, but we consider only the top gate bias in the following discussion[8]. The energy band diagram and the k-space occupation of the 2D electron modes are shown in Fig. 19.5 (d) for $V_{ds} = 0$, which is identical to the last section. Parts (e) and (f) of the figure show how the k-space occupation function changes at the source-injection point in response to the drain voltage. The source and drain Fermi levels E_{Fs} and E_{Fd} are related by the applied voltage $E_{Fs} - E_{Fd} = qV_{ds}$. The carriers injected from the source and drain contacts are indicated with different colors for the purpose of illustration.

Since the drain is now biased, the local mobile electron concentration $n_s(x)$ must change under the gate in the channel from the source to the drain. We will account for this spatial variation in later sections, when we introduce scattering in the channel. However, for the ballistic FET, the device characteristics can be obtained by investigating the affairs at the source injection point, which is the region near the source where the slope of the conduction band edge is zero, called

---

[6] Since the electron sheet density n_s is always a positive number, it is apparent that the parallel-plate capacitor approximation $n_s \approx \frac{C_b C_q}{C_b + C_q} (V_{gs} - V_T)$ cannot be correct for $V_{gs} < V_T$.

[7] The distinct temperature dependence below and above threshold: insulator for $V_{gs} < V_T$, and metallic for $V_{gs} > V_T$ can be viewed as a gate-controlled **metal-insulator transition**: this is a characteristic of every FET. By connecting with several earlier chapters, especially the foundational material of Chapter 5, the general relationship of the mobile carrier density with source and drain quasi-Fermi levels for FET channels of d-dimensions is given by $n_d = \frac{1}{2} \cdot N_c^d \cdot [F_{\frac{d-2}{2}}(\eta_s) + F_{\frac{d-2}{2}}(\eta_d)]$, where $F_j(\eta)$ is the Fermi–Dirac integral of order j, $\eta_s = \frac{E_{Fs} - E_c}{k_b T}$, and $\eta_d = \frac{E_{Fd} - E_c}{k_b T}$.

[8] The double-gate case is in Exercise 19.4.

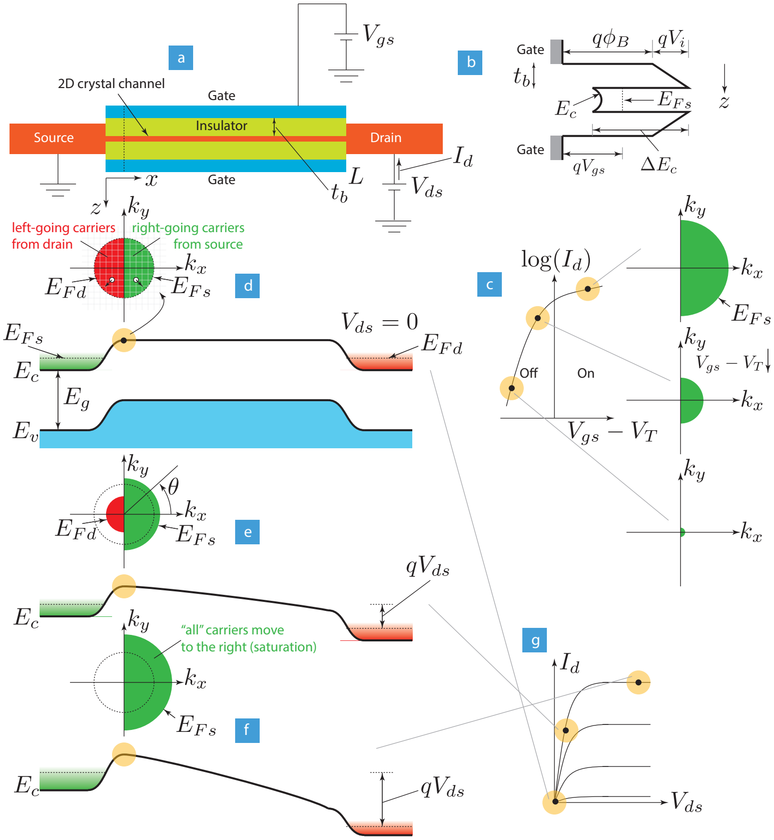


Fig. 19.5 A ballistic FET composed of a thin, 2D semiconductor channel is shown in (a), with the energy band diagram in the vertical z-direction shown in (b). The current that flows in the semiconductor channel I_d is shown in (c) in response to the gate voltage V_{gs} at a fixed drain voltage V_{ds} in the log scale. Note the similarity to Fig. 19.4. The energy band diagram in (d) shows the case for $V_{ds} = 0$, and (e, f) are for drain voltages $V_{ds} > 0$. The drain current I_d in response to the drain voltage V_{ds} is shown in (g), each curve corresponding to a different gate voltage V_{gs}. The drain current increases as a gate-controlled resistor for small V_{ds}, and saturates at high V_{ds}. The current-voltage characteristics hold for both ballistic and scattering dominated FETs. The k-space occupation functions depicted are only for the ballistic case, at the source injection point, which is the location where the conduction band reaches its maximum. This source injection point is also called the "top of the barrier" in transistor device physics.

the "top of the barrier". Imagine being perched at this point, and counting electrons whizzing to the right and to the left. Because the transport is ballistic and there is no scattering, electrons that move to the right with $k_x > 0$ must have begun their journey from the source contact. This includes electrons with non-zero k_y: they are moving at an angle $\theta = \tan^{-1}(k_y/k_x)$ to the x-axis. All right-going electrons are in equilibrium with the source electrode, and hence their population is characterized by a quasi-Fermi level equal to the source: E_{Fs}. The electrons at this point moving to the left share the quasi-Fermi level of the drain. But because $V_{ds} > 0$, it has become energetically much more difficult for the drain to inject carriers that make it to the source injection point, since it is at an energy qV_{ds} below the source Fermi level. Electrons injected from the drain that have energy lower than the top of the barrier are reflected back to the drain. Thus, there are much fewer electrons moving to the left at the source injection point, as indicated by a smaller left-going occupied state semi-circle compared to the larger right-going half circle[9] as seen in Fig. 19.5 (e).

The total area of the two semi-circles is proportional to the local 2D sheet electron density at the source injection point. At $V_{ds} = 0$, the den-

[9] At a high drain voltage, the source loses electrons and the drain gains them. To maintain charge neutrality, $E_{Fs} - E_c$ increases to compensate for the lost electrons, and $E_{Fd} - E_c$ decreases to account for those gained: this is shown schematically in Fig. 19.5 (f). Note that here E_c is the conduction-band edge at the source injection point. Because the source loses high-energy electrons, it undergoes evaporative cooling, whereas the drain is heated since the ballistic electrons lose their energy by relaxing to the Fermi level after entering the drain. Note that the **drain current saturates** at high drain voltages: this is because only the right-going semi-circle of occupied states survive, and the shape of occupied states does not change beyond a drain voltage. The saturated drain current only changes with gate voltage which controls the area of the right-going semi-circle of occupied states, as shown in Figures 19.5 (g, c, and f).

sity $n_s(V_{gs})$ depends only on the gate voltage, and is given by Equation 19.2 everywhere in the channel. How does the area change with the drain voltage at the source injection point? **For a well-designed FET, to an excellent approximation, the area does not change: it depends only on the gate voltage, and not on the drain voltage.**[10] As a consequence, the right-half circle must increase in area, and the left-half circle must shrink, in a manner that the area remains constant - as shown in Figures 19.5 (e) and (f). The quasi Fermi levels E_{Fs} and E_{Fd} must adjust such that the drain-control relation $E_{Fs} - E_{Fd} = qV_{ds}$ is met *simultaneously* for the **drain control** Equation 19.1, and Equation 19.2, which is the **gate control** equation. Defining $qV_{ds}/k_bT = v_d$, $E_{Fs} - E_{Fd} = qV_{ds}$ is rewritten as $\eta_s - \eta_d = v_d$. The gate control equation yields a unique 2D sheet density corresponding to the gate voltage denoted by $n_s(V_{gs})$. Rearranging the drain control equation reveals that it is a quadratic equation in disguise, with a solution:

$$n_s(V_{gs}) = \frac{n_q}{2}(\ln[1+e^{\eta_s}] + \ln[1+e^{\eta_d}])$$

$$\implies (1+e^{\eta_s})(1+e^{\eta_s - v_d}) = e^{\frac{2n_s(V_{gs})}{n_q}}$$

$$\implies \boxed{\eta_s = \ln[\frac{\sqrt{(1+e^{v_d})^2 + 4e^{v_d}(e^{\frac{2n_s(V_{gs})}{n_q}} - 1)} - (1+e^{v_d})}{2}].} \quad (19.4)$$

The boxed solution in Equation 19.4 is the quantitative dependence of the source quasi Fermi level $\eta_s = (E_{Fs} - E_c)/k_bT$ on the gate voltage through the relation $n_s(V_{gs})$ in Equation 19.2, and the drain voltage through v_d. The quantum current is now obtained by summing the currents carried by each occupied (k_x, k_y) state:[11]

$$I_d = \frac{qg_sg_v}{L}\sum[\mathbf{v}_g(k_x,k_y)\cdot\hat{x}]\cdot f(k_x,k_y)$$

$$= \frac{qg_sg_v}{L}\int\frac{dk_y}{\frac{2\pi}{W}}\cdot\frac{dk_x}{\frac{2\pi}{L}}\cdot\underbrace{[\mathbf{v}_g(k_x,k_y)\cdot\hat{x}]}_{\hbar k_x/m_c^\star}\cdot f(k_x,k_y)$$

$$\implies \frac{I_d}{W} = \frac{qg_sg_v\hbar}{(2\pi)^2m_c^\star}\int_0^\infty k\cdot dk\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}}d\theta\cdot\frac{k\cos\theta}{1+\exp[\frac{E_c+\frac{\hbar^2k^2}{2m_c^\star}-E_{Fs}}{k_bT}]} -$$

$$\frac{qg_sg_v\hbar}{(2\pi)^2m_c^\star}\int_0^\infty k\cdot dk\int_{+\frac{\pi}{2}}^{+\frac{3\pi}{2}}d\theta\cdot\frac{k\cos\theta}{1+\exp[\frac{E_c+\frac{\hbar^2k^2}{2m_c^\star}-E_{Fd}}{k_bT}]}$$

$$\implies \boxed{\frac{I_d}{W} = \underbrace{\frac{q^2}{h}\cdot N_c^{1d}\cdot\frac{k_bT}{q}}_{J_0^{2d}}\cdot[F_{\frac{1}{2}}(\eta_s) - F_{\frac{1}{2}}(\eta_s - v_d)],} \quad (19.5)$$

where $N_c^{1d} = g_sg_v(\frac{2\pi m_c^\star k_bT}{h^2})^{\frac{1}{2}}$ is a 1D band-edge DOS.

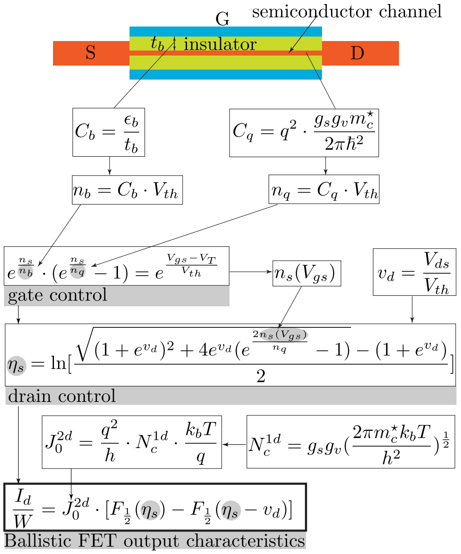

G
semiconductor channel
S
t_b
insulator
D
$C_b = \frac{\epsilon_b}{t_b}$
$C_q = q^2\cdot\frac{g_sg_vm_c^\star}{2\pi\hbar^2}$
$n_b = C_b\cdot V_{th}$
$n_q = C_q\cdot V_{th}$
$e^{\frac{n_s}{n_b}}\cdot(e^{\frac{n_s}{n_q}} - 1) = e^{\frac{V_{gs}-V_T}{V_{th}}}$
gate control
$n_s(V_{gs})$
$v_d = \frac{V_{ds}}{V_{th}}$
$\eta_s = \ln[\frac{\sqrt{(1+e^{v_d})^2 + 4e^{v_d}(e^{\frac{2n_s(V_{gs})}{n_q}} - 1)} - (1+e^{v_d})}{2}]$
drain control
$J_0^{2d} = \frac{q^2}{h}\cdot N_c^{1d}\cdot\frac{k_bT}{q}$
$N_c^{1d} = g_sg_v(\frac{2\pi m_c^\star k_bT}{h^2})^{\frac{1}{2}}$
$\frac{I_d}{W} = J_0^{2d}\cdot[F_{\frac{1}{2}}(\eta_s) - F_{\frac{1}{2}}(\eta_s - v_d)]$
Ballistic FET output characteristics

Fig. 19.6 Ballistic FET flowchart.

---

[10] This approximation is the key to quantitatively obtaining the output characteristics of the ballistic FET in a simple way. The approximation can be removed to allow for a capacitive coupling between the drain and the source injection point, but we avoid it in the treatment here.

[11] For a parabolic bandstructure, the group velocity projected along the x-direction is $v_{gx} = \hbar k_x/m_c^\star$. Using radial coordinates, $(k_x, k_y) = k(\cos\theta, \sin\theta)$.

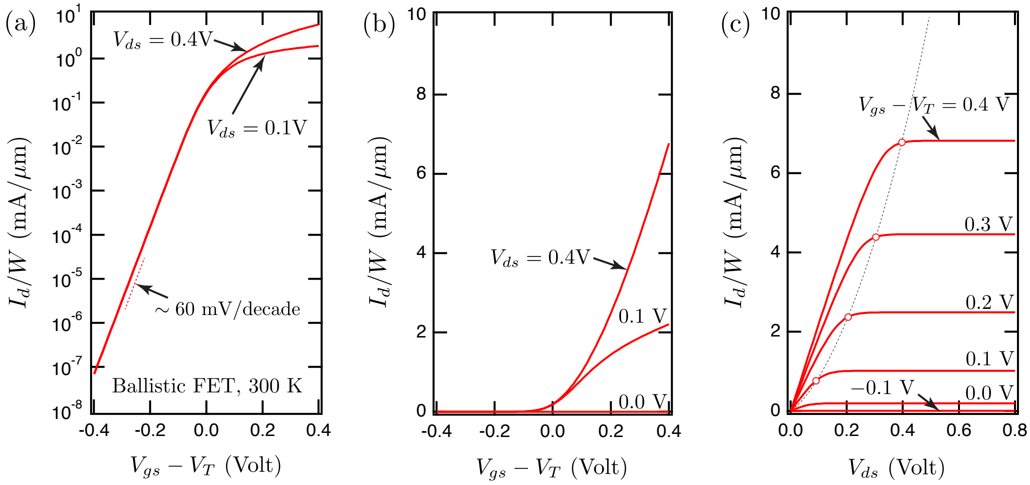


Fig. 19.7 The current-voltage characteristics of a ballistic FET at 300 K with $\epsilon_b = 20\epsilon_0$ and $t_b = 2$ nm. The semiconductor channel has a conduction band edge electron effective mass $m_c^\star = 0.2m_e$, and a conduction band valley degeneracy $g_v = 1$. The band properties are of the wide-bandgap semiconductor GaN. The effective mass is similar to silicon, which has larger g_v, as discussed later.

The above relations, summarized in the flowchart in Fig. 19.6, provide the complete ballistic FET characteristics of interest. The ballistic FET current written as $I_d/W = J_0^{2d} \cdot [F_{\frac{1}{2}}(\eta_s) - F_{\frac{1}{2}}(\eta_s - v_d)] = J_R - J_L$, where $J_R = J_0^{2d} \cdot F_{\frac{1}{2}}(\eta_s)$ and $J_L = J_0^{2d} \cdot F_{\frac{1}{2}}(\eta_s - v_d)$, shows the net current as the difference of the right-going and the left-going currents[12].

19.3 Ballistic I-V characteristics

Fig. 19.7 shows the output characteristics of a ballistic FET calculated using the formalism developed in the last section. The procedure for the calculation is summarized in the flowchart shown in Fig. 19.6 as three primary equations, that of the **gate control**, **drain control**, and the **ballistic FET output** current as a function of the gate and drain voltages. The only non-analytical step is the transcendental gate control equation that relates the 2D sheet density $n_s(V_{gs})$ with the gate voltage, which is easily solved on a computer. By discussing a few features of the FET characteristics, we will recognize analytical approximations of the off-state and on-state characteristics, which will significantly aid our understanding without the need of the computer.

Fig. 19.7 (a) and (b) show the drain current per unit width as a function of the excess gate voltage $V_{gs} - V_T$ in the log and linear scales respectively, for two values of drain voltages: $V_{ds} = 0.1$ V and $V_{ds} = 0.4$ V. The current for gate voltages below threshold ($V_{gs} < V_T$) decreases exponentially with the gate voltage. It has a weak dependence on the drain voltage, and is called the off-state of the FET, characterized by an **exponential dependence** on the gate voltage. The current for gate voltages $V_{gs} > V_T$ is the on-state, which is characterized by a **polynomial dependence** on the gate voltage. Fig. 19.7 (c) shows the drain current per unit width as a function of the drain voltage for various V_{gs}. It is characterized by a roughly **linear regime** for $V_{ds} < V_{gs} - V_T$, and a **saturated current** for $V_{ds} > V_{gs} - V_T$, with a smooth transition across the $V_{ds} \approx V_{gs} - V_T$ contour, shown as a dashed line. This is a general characteristic for all FETs. Let us investigate the limiting regimes, for which we can obtain analytical forms.

---

[12] Note that in Equation 19.5, the gate length L from the source to drain does not appear. This is a characteristic of ballistic, quantum transport of electrons that do not suffer scattering in their motion in the channel. At $T = 300$ K, for $m_c^\star = 0.2m_e$ and $g_v = 1$, the coefficient of the ballistic current density is $J_0^{2d} \sim 0.2$ mA/μm, a useful number to remember. Note also the current density coefficient J_0^{2d} is in the form of the product of the quantum of conductance, the thermal voltage, and the band-edge DOS of a dimension *lower* than the channel carriers. The generalization of the ballistic current for d-dimensions is $J_d = \frac{q^2}{h} \cdot N_c^{d-1} \cdot \frac{k_b T}{q} \cdot [F_{\frac{d-1}{2}}(\eta_s) - F_{\frac{d-1}{2}}(\eta_s - v_d)]$. Expressions for all dimensions are listed in Fig. A.1 of the Appendix for easy reference.

Off-state: The **off-state** of the FET was discussed in Equation 19.3; the solution to the gate-control equation is $n_s \approx n_q \exp{[(V_{gs} - V_T)/V_{th}]}$, and since $\eta_s << -1$ for a low carrier density at the source injection point, a few algebraic steps simplifies the drain control equation, and the FET current equation to[13]

$$e^{\eta_s} \approx e^{\frac{V_{gs}-V_T}{V_{th}}} \cdot \frac{1}{1+e^{-v_d}} \implies \boxed{\frac{I_d}{W}|_{off} \approx J_0^{2d} \cdot e^{\frac{V_{gs}-V_T}{V_{th}}} \cdot 2\tanh{[\frac{V_{ds}}{2V_{th}}]}.} \tag{19.6}$$

On-state: For gate voltages $V_{gs} - V_T >> V_{th}$, the FET is in the **on-state**. The ballistic current in this limit is[14]

$$\boxed{\frac{I_d}{W}|_{on} \approx J_0^{2d} \cdot [\frac{\eta_s^{\frac{3}{2}}}{\Gamma(2+\frac{1}{2})}] \approx J_0^{2d} \cdot \frac{4}{3\sqrt{\pi}} \cdot (\frac{2C_g(V_{gs} - V_T)}{qn_q})^{\frac{3}{2}}}$$

$$\implies \frac{I_d}{W}|_{on} \approx \frac{16}{3} \cdot \frac{qg_s g_v\sqrt{m_c^\star}}{h^2} \cdot [\frac{q(V_{gs} - V_T)}{1 + \frac{q^2 \cdot \frac{g_s g_v m_c^\star}{2\pi\hbar^2}}{\frac{\epsilon_b}{t_b}}}]^{\frac{3}{2}}. \tag{19.7}$$

To obtain the on-state current expression, we use the degenerate approximation of Fermi–Dirac integrals to write $\eta_s = \frac{2C_g(V_{gs}-V_T)}{qn_q}$. Since $\eta_s >> +1$, the Fermi–Dirac integral in the expression for the ballistic current approximates to its degenerate, or polynomial limit $F_{1/2}(\eta) \approx \eta^{3/2}/\Gamma(5/2)$, yielding the on-state ballistic current limit of Equation 19.7. Since I_d is saturated, V_{ds} does not appear explicitly.

Performance: The highest on-current maximizes the FET performance. The on-state ballistic current has a non-monotonic dependence on the semiconductor band-edge effective mass $m_c^\star$ and valley degeneracy g_v because of their appearance in both the numerator and the denominator of Equation 19.7. The maxima, inferred by taking a derivative of Equation 19.7, occurs when $C_q \sim C_b$, i.e. when the semiconductor quantum capacitance is nearly matched to the barrier capacitance. As shown in Fig. 19.8, for a given C_b, the current is maximized for a semiconductor that meets specific $m_c^\star$ and g_v. Fig. 19.8 indicates that for the chosen barrier, the ideal semiconductor channel with $g_v = 1$ should have $m_c^\star \sim 0.15 m_e$. For large effective masses, a lower valley degeneracy is desirable since $\frac{I_d}{W}|_{on} \sim \frac{1}{m_c^\star\sqrt{g_v}}$, whereas for low effective masses, a larger valley degeneracy is desired, since $\frac{I_d}{W}|_{on} \sim g_v\sqrt{m_c^\star}$.

Effective mass, valley, and spin degeneracies: Fig. 19.9 shows the effect of the band parameters of the semiconductor channel on the

[13] We have used the Maxwell–Boltzmann approximation of the Fermi–Dirac integrals in the equation for the current, since for $\eta << -1$, $F_j(\eta) \approx e^{\eta}$. The off-state current becomes independent of V_{ds} for $V_{ds} >> V_{th}$, since $\tanh{(x \to \infty)} = 1$. The exponential dependence of the off-state drain current on the gate voltage is thus seen explicitly.

[14] Note that in the on state, $qn_s \approx C_g(V_{gs} - V_T)$, where $C_g = C_b C_q/(C_b + C_q)$. The quasi Fermi level at the source injection point for the right-going carriers is $\eta_s >> +1$ since the gate has populated the channel with a lot of electrons. Let us consider the case when the drain voltage is $V_{ds} >> V_{gs} - V_T$, implying $\eta_s - v_d << -1$. Under this situation, the sheet density at the source injection point is $n_s \approx \frac{n_q}{2}\eta_s$, implying $\eta_s \approx 2n_s/n_q$.

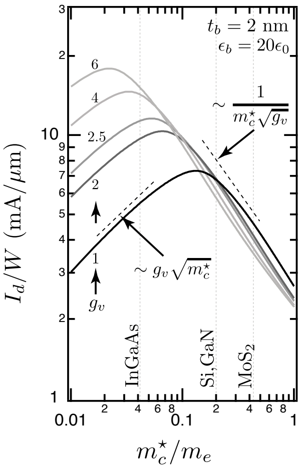


Fig. 19.8 The calculated current in a ballistic FET at 300 K for semiconductor channels with varying valley degeneracies and band edge effective masses for $V_{gs} - V_T = 0.4$ V and $V_{ds} = 0.4$ V for barrier thickness $t_b = 2$ nm and dielectric constant $\epsilon_b = 20\epsilon_0$. At current saturation, the maximum current is achieved when the barrier capacitance and the quantum capacitance are nearly matched: $C_b \approx C_q$. The actual current drives are typically lower from the calculated ballistic values due to contact resistances.

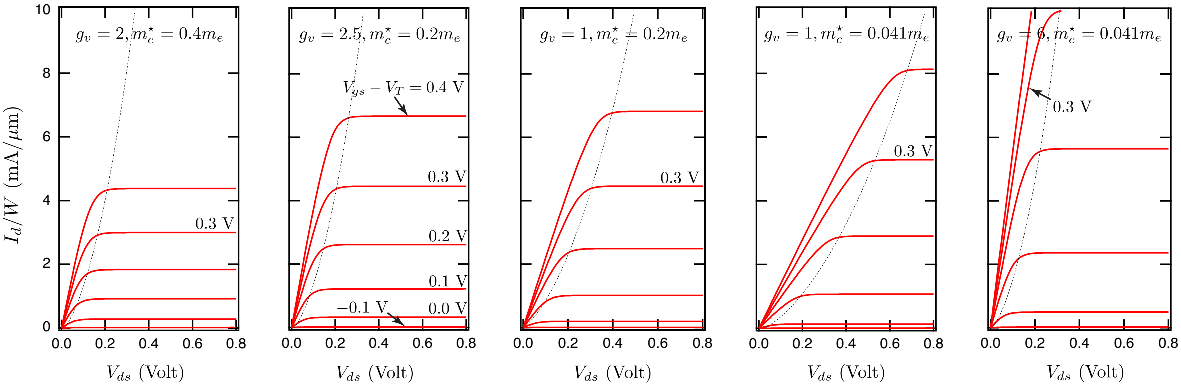


Fig. 19.9 Effect of the semiconductor channel band parameters on ballistic FET $I_d - V_{ds}$ characteristics for a gate barrier of dielectric constant $\epsilon_b = 20$ and thickness $t_b = 2$ nm at 300 K. More valleys reduce the drain voltage required to obtain current saturation.

characteristics of the ballistic FET for a thin, high-κ dielectric gate insulator. For semiconductors with small effective masses, a large valley degeneracy is seen to reduce the V_{ds} required to achieve current saturation. In discussions on semiconductor bandstructure, especially in Chapter 12 on the $\mathbf{k} \cdot \mathbf{p}$ theory, we have seen that small effective masses at the band edges are related to small energy bandgaps. For sp^3 bonded direct-bandgap semiconductors, we found a crude relationship of $m_c^\star \approx E_g/20$, where E_g is in eV. Though small effective mass semiconductors can provide high on-currents, it is typically difficult to turn them off effectively, since the Fermi level can move easily through the small bandgap to the undesired band, for example creating holes in a n-channel device. This leads to effects such as interband tunneling and impact ionization at high drain voltages, which make it difficult to obtain very large on-off ratios for FETs formed of very small bandgap and small-effective mass semiconductors[15].

For most semiconductors, the spin degeneracy is $g_s = 2$ for each energy eigenvalue $E(k)$. This implies the total number of electrons that can be fit into any k-state is $g_s g_v = 2g_v$. There are some notable exceptions. For example, in some 2D semiconductor crystals such as monolayer WSe$_2$, there are $g_v = 2$ inequivalent conduction and valence band valleys in the first Brillouin zone, called the K and K' valleys. Because of the crystal symmetries, the spin-states are split differently in different valleys. This is shown in Fig. 19.10. The valence bands in the K and K' valleys are split into two spin-polarized bands. If a p-channel FET is made with such a semiconductor, then the correct factor of $g_s g_v$ must be used to obtain the ballistic current density. For example, if the Fermi level lies *between* the two spin-split valence bands, the spin-valley degeneracy factor is $g_s g_v = 1 \cdot g_v = 2$, and not 4.

Injection velocity: The ballistic FET output characteristics that was shown in Fig. 19.7 (c) is re-plotted in the left of Fig. 19.11 to indicate three distinct phases of operation. For a drain voltage of $V_{ds} = 0.15$ V,

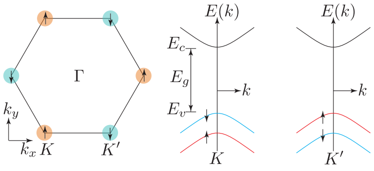


Fig. 19.10 Spin-valley splitting in the valence bands of transition-metal dichalcogenide 2D semiconductors such as WSe$_2$. If the Fermi level is between the 2 valence bands, there are spin-polarized hole pockets in the k-space, in which each state has a spin-degeneracy of $g_s = 1$, and not 2. Spin-splitting occurs in conventional semiconductors in a magnetic field, and in the absence of external magnetic fields if the semiconductor is intrinsically magnetic, such as in dilute magnetic semiconductors.

---

[15] A limiting case is obtained for 2D graphene, which has a zero bandgap. Though the bandstructure of 2D graphene is linear, and not parabolic, evaluation of the ballistic characteristics may be obtained in a similar fashion as for the parabolic bands. Since the gap is zero, the on/off ratio of a FET made with a graphene channel is very small.

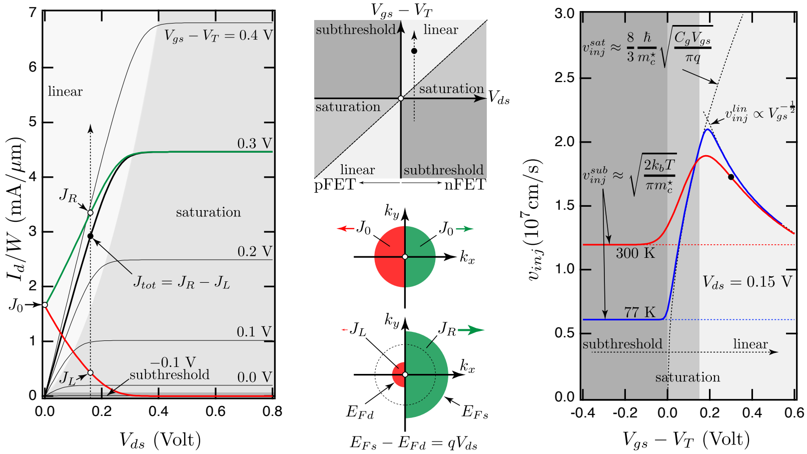


Fig. 19.11 Left: Output characteristics of a ballistic FET from Fig. 19.7 (c) showing the phases of operation, the right-going ballistic current J_R, and the left-going component J_L at $V_{gs} - V_T = 0.3$ V; the net current is $J_{tot} = J_R - J_L$. The k-space occupation functions at this gate voltage and $V_{ds} = 0.15$ V are shown in the center, bottom plot. Center, top: Phase diagram of operation of the ballistic FET as a function of the gate and drain voltages for nFETs and pFETs, showing the subthreshold, saturation, and linear regimes. Right: The injection velocity v_{inj} at the source injection point of the ballistic FET as a function of $V_{gs} - V_T$ at two temperatures calculated using Equation 19.8, demarcating the three phases of operation.

the gate sweeps the device from the **subthreshold** region, through a **saturation** regime, into a **linear** regime along the dashed line. The corresponding phase diagram is shown in the middle for the nFET, and also a corresponding pFET as a function of the drain and gate voltages[16]. When the drain voltage is increased, more right-going states are occupied, while not changing the total occupied area as shown in the k-space. This increases J_R and reduces J_L as seen in the output characteristics. J_L is a reasonable fraction of the net current in the linear regime of operation. When V_{ds} exceeds roughly $V_{gs} - V_T$, J_L becomes negligible, most of the current is by right-going states J_R, and the current saturates. To connect with the classical notion of current being a product of charge and velocity, one can define an ensemble injection velocity v_{inj} as simply the net current divided by the net sheet charge qn_{inj} at the source injection point via the relation[17]

$$v_{inj} = \frac{I_d/W}{qn_{inj}} = \frac{\frac{q^2}{h}\cdot N_c^{1d}\cdot\frac{k_bT}{q}\cdot[F_{\frac{1}{2}}(\eta_s) - F_{\frac{1}{2}}(\eta_s - v_d)]}{\frac{q\cdot n_q}{2}\cdot[F_0(\eta_s) + F_0(\eta_s - v_d)]}$$

$$\implies \boxed{v_{inj}^{sub} \approx \sqrt{\frac{2}{\pi}\cdot\frac{k_bT}{m_c^\star}}\cdot\tanh(\frac{qV_{ds}}{2k_bT})} \text{ for } V_{gs} < V_T$$

$$\implies \boxed{v_{inj}^{sat} \approx \frac{\hbar}{m_c^\star}\sqrt{\frac{128}{9\pi g_s g_v}\cdot\frac{C_g\cdot(V_{gs} - V_T)}{q}}} \text{ for } V_{gs} > V_T. \qquad (19.8)$$

Fig. 19.11 shows v_{inj} on the right for the ballistic FET of Fig. 19.7 for two temperatures[18], seen to be in the range of 10^7 cm/s.

Surface potential and gate control: The electrostatic control of the mobile carrier density at the source-injection point by the voltage V_{gs} applied on the gate is critical to the switching efficiency, gain, and speed of the transistor. The energy band diagram at the top of Fig.

[16] The k-space occupation functions at the source injection point show that for a fixed gate voltage, at $V_{ds} = 0$ V equal number of right-going and left-going states are occupied, leading to equal currents $J_R = J_L = J_0$, and net zero current. This current for $V_{gs} - V_T = 0.3$ V is seen in the left figure to be $J_0 \approx 1.6$ mA/μm.

[17] The approximation for the subthreshold regime is obtained using $\eta_s << -1$ and $F_j(\eta) \approx e^{\eta}$, and that for the saturation regime for $\eta_s >> +1$ when $F_j(\eta) \approx \eta^{3/2}/\Gamma(5/2)$: notice how often these approximations appear!

[18] The subthreshold v_{inj} is entirely thermal, independent of the gate voltage, and also independent of the drain voltage for $V_{ds} >> V_{th}$. On the other hand, v_{inj} in the saturation and linear regimes are only weakly temperature-dependent, and controlled mostly by the gate voltage: its dependence on V_{gs} can be concluded from the k-space occupation of semi-circles. The injection velocity is useful because if the gate-controlled n_{inj} is known, the current can be obtained in a simple fashion. However, since it is a weighted mean, one must exercise caution in attaching too much physical meaning to it. It is clear from the k-space image that there is a large spread in electron velocity vectors, some of them are actually going from the drain → source, and some at right angles, along the width of the FET! Exercise 19.6 discusses injection velocity further.

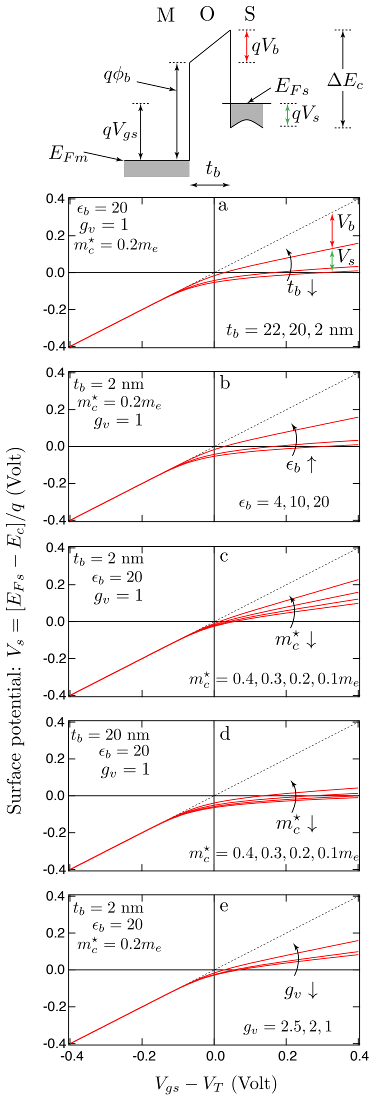


Fig. 19.12 The dependence of the surface potential of the semiconductor channel in a FET on various transistor design parameters at 300 K.

19.12 shows how the voltage applied on the gate divides into two parts: a voltage drop V_b in the (oxide) barrier and a voltage drop V_s in the semiconductor, that goes into creating the mobile carriers in the channel and is typically referred to as the *surface potential*. It is desirable to have a significant fraction go into creating the mobile carriers that actually change the conductivity of the channel, whereas V_b is the portion that sustains the electric field-effect in the barrier which avoids carriers from flowing from the channel into the gate. Let us investigate how the choices of the barrier materials via ϵ_b and t_b, and the semiconductor channel material via $m_c^\star$ and g_v affect exactly how this voltage division occurs.

Let $q\phi_b$ be the M-O barrier height, and ΔE_c the O-S conduction band offset. The field in the barrier is $F_b = \frac{V_b}{t_b} = \frac{qn_s}{\epsilon_b}$. By following the energy in the loop from the metal Fermi level E_{Fm} to the semiconductor Fermi level E_{Fs} in a manner similar to Chapter 17, Section 17.3, we obtain for a sheet density n_s in the semiconductor channel,

$$q\phi_b + q(\frac{qn_s}{\epsilon_b})t_b - \Delta E_c + (E_{Fs} - E_c) = qV_{gs}$$

$$qV_T = (q\phi_b - \Delta E_c) \implies (V_{gs} - V_T) = \underbrace{q\frac{n_s}{C_b}}_{V_b} + \underbrace{\frac{E_{Fs} - E_c}{q}}_{V_s}, \qquad (19.9)$$

which shows the voltage division explicitly. By using the exact relation of n_s with V_{gs} from Equation 19.2, the voltage division is quantitatively obtained, as shown in Fig. 19.12 (a-e) for several choices of the barrier and semiconductor channel. Panel (a) shows the division of voltage against the dashed line which is the $V_s = V_{gs} - V_T$ limit. We note that the entire gate voltage does transfer to the semiconductor channel under subthreshold conditions when $V_{gs} << V_T$. This is because there are very few mobile carriers in the semiconductor channel since the semiconductor Fermi level $E_{Fs} < E_c$ is inside the bandgap, and the net field in the oxide is negligible. As V_{gs} approaches V_T and exceeds it, the mobile charge density in the semiconductor increases, increasing the electric field in the barrier, and the voltage starts to divide. From panel (a), which is for a high-κ dielectric with $\epsilon_b = 20\epsilon_0$ and a semiconductor channel with $m_c^\star = 0.2m_e$ and $g_v = 1$, the surface potential is the largest for the thinnest barrier with $t_b = 2$ nm. For this barrier thickness, at $V_{gs} - V_T = +0.4$ V, the surface potential reached is $V_s = (E_{Fs} - E_c)/q \approx 0.15$ V, the semiconductor Fermi level E_{Fs} is 150 meV above the lowest conduction band energy E_c, which implies that $\eta_s = 150/26 >> +1$, and the carrier density is degenerate. The voltage dropped in the barrier at this gate voltage is $V_b = 0.4 - 0.15 = 0.25$ V. As t_b is increased from 2 nm to 20 and 22 nm, V_s drops and V_b increases, since a thick barrier implies a larger voltage drop in the oxide, and consequently weaker control of the channel conductivity by the gate. Note that for a thick barrier, the carrier density is only weakly degenerate since $V_s > 0$ only for the highest V_{gs}. A thinner barrier is desirable for the best gate control, as long as the barrier can prevent carrier flow between the gate metal and the semiconductor channel.

The dependence of the barrier material dielectric constant ϵ_b is shown in panel (b), a higher value is desirable. Panel (c) shows the dependence of the band-edge effective mass of the semiconductor channel for a thin high-κ dielectric: a smaller effective mass implies a smaller density of states, which requires a larger V_s for the same carrier density. But for a thicker barrier of panel (d), the dependence on the effective mass is weaker since most of the voltage drops in the barrier. Panel (e) shows the effect of the valley degeneracy of the semiconductor channel for a thin, high-κ dielectric: a smaller valley degeneracy leads to a higher surface potential because of a smaller DOS[19].

19.4 Quantum wire ballistic FET

In the last section, the ballistic FET characteristics were discussed for a semiconductor that had a 2-dimensional channel. The need for stronger gate control over the semiconductor channel, and for miniaturization and denser packing makes FETs in which the semiconductor channel is one-dimensional attractive. Fig. 19.13 shows a 1D quantum-wire (or nanowire) FET. The quantum wire semiconductor channel is shown as a long cylinder[20] of radius r_w, around which an insulating (oxide) barrier of thickness t_b is wrapped, followed by a gate metal that is wrapped around the oxide. The source and drain ohmic contacts are made to the ends of the wire. We have already identified in Chapter 5 the ballistic current for a d-dimensional channel to be $J_d = \frac{q^2}{h} \cdot N_c^{d-1} \cdot V_{th} \cdot [F_{\frac{d-1}{2}}(\eta_s) - F_{\frac{d-1}{2}}(\eta_s - v_d)]$, and the gate/drain control equation of the charge at the source injection point as $n_d = \frac{N_c^d}{2}[F_{\frac{d-2}{2}}(\eta_s) + F_{\frac{d-2}{2}}(\eta_s - v_d)]$. If we can relate the applied voltages V_{gs} and V_{ds} to the η_s, we can obtain the characteristics of the ballistic 1D FET using $d = 1$.

The 1D barrier capacitance for the cylindrical FET structure is $C_b = 2\pi\epsilon_b / \ln{(1 + \frac{t_b}{r_w})}$ in F/cm units. To obtain this apply Gauss's law $\oint \mathbf{D} \cdot d\mathbf{S} = Q_{encl}$ inside the barrier layer at the source injection point[21]. For a density of mobile 1D carriers in the channel of n_{1d}, for a length L Gauss's law requires $\epsilon_b \cdot E(\rho) \cdot 2\pi\rho \cdot L = q \cdot n_{1d} \cdot L$, which yields the electric field profile $E(\rho) = qn_{1d}/(2\pi\epsilon_b\rho)$, a well-known result in electrostatics for a cylindrical capacitor. Note that unlike the 2D FET in which the electric field was constant in the oxide, for the quantum wire the **radial electric field increases from the metal to the semiconductor**: inside the oxide it is higher at the semiconductor-oxide interface than the oxide-metal interface. The line-integral of the electric field in the oxide layer in the radial direction gives the voltage drop $V_b = -\int_{\rho=r_w}^{\rho=r_w+t_b} d\rho \cdot E(\rho) = \frac{qn_{1d}}{2\pi\epsilon_b} \cdot \ln{(\frac{r_w+t_b}{r_w})} = \frac{qn_{1d}}{C_b}$, where $C_b = 2\pi\epsilon_b / \ln{(1 + \frac{t_b}{r_w})}$ is the barrier capacitance per unit length. The voltage division between the barrier and the semiconductor thus leads to the complete 1D ballistic FET characteristics:

---

[19] The discussion here has been limited to parabolic bandstructure, but may be re-done for crystals with linear (conical) bandstructure, or other cases. The same theoretical formalism that applies to gapped semiconductors applies to zero-gap graphene, metals, normal and Weyl semimetals, and topological insulators: all that is necessary is the correct bandstructure $E(k)$ be used with the corresponding distribution functions. For gapped semiconductors that have a parabolic band-edge dispersion, if the Fermi level samples regions of the bandstructure that have significant non-parabolicity, minor changes to the model are necessary: instead of $m_c^\star$, the actual bandstructure $E(k)$ must be used to re-derive the expressions for electrostatics and ballistic currents.

[20] Quantum wire FETs with non-circular cross-sections and multiple subbands are are also possible and may be treated numerically.

[21] The displacement vector is $\mathbf{D} = \epsilon_b E(\rho)\hat{\rho}$, where ρ is the radial distance from the center of the quantum wire, $\hat{\rho}$ is the radial unit vector, $E(\rho)$ is the radial electric field, and Q_{encl} is the charge enclosed in the Gaussian volume.

$$\underbrace{\frac{V_{gs}-V_T}{V_{th}}}_{v_g} = \underbrace{\frac{q\cdot n_{1d}}{C_b V_{th}}}_{n_{1d}/n_b^{1d}} + \underbrace{\frac{E_{Fs}-E_c}{qV_{th}}}_{\eta_s}$$

$$n_{1d} = \frac{N_c^{1d}}{2}[F_{-\frac{1}{2}}(\eta_s) + F_{-\frac{1}{2}}(\eta_s - v_d)]$$

$$\implies \boxed{n_{1d} = \frac{N_c^{1d}}{2}[F_{-\frac{1}{2}}(v_g - \frac{n_{1d}}{n_b^{1d}}) + F_{-\frac{1}{2}}(v_g - \frac{n_{1d}}{n_b^{1d}} - v_d)]}$$

$$\boxed{I_{1d} = \frac{q^2}{h}\cdot g_s g_v \cdot \frac{k_b T}{q}\cdot \ln(\frac{1+e^{\eta_s}}{1+e^{\eta_s - v_d}})}$$

$$\underbrace{\ln(\frac{1+e^{\eta_s}}{1+e^{\eta_s - v_d}}) \approx [\eta_s - (\eta_s - v_d)]}_{\eta_s >> +1,\text{ and } v_d << \eta_s} \implies I_d \approx [\underbrace{\frac{q^2}{h}\cdot g_s g_v}_{G_Q}]\cdot V_{ds}. \quad (19.10)$$

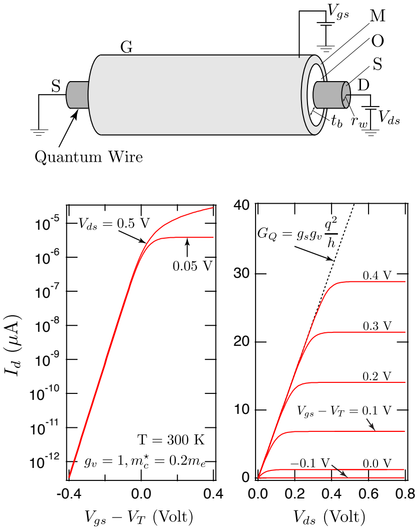


Fig. 19.13 A ballistic quantum wire FET in which the transport is one-dimensional. The calculation is for a quantum wire radius $r_w = 1$ nm and a insulating barrier thickness $t_b = 2$ nm with a dielectric constant $\epsilon_b = 20$. Note that even at 300K, the I_d/V_{ds} in the linear regime is a quantized conductance, $G_Q = 2q^2/h \sim [12950\Omega]^{-1}$.

Fig. 19.13 shows the output characteristics calculated using Equation 19.10 for a quantum wire ballistic FET[22]. The barrier and semiconductor material properties and geometries are shown in the figure. The transfer characteristics I_d vs. V_{gs} show several orders of switching, similar to the 2D FET. In the on-state, the output characteristics seen on the right indicate saturation current in the 10s of μA; this is typical of ballistic (or near-ballistic) 1D channel FETs with a single allowed mode. Comparing the output characteristics to the 2D case in Fig. 19.7 (c), a clear difference is seen in the linear regime. **Whereas the output characteristics of a 2D FET in the linear regime is that of a gate-controlled resistor, that of the 1D ballistic FET is not. The I_d curves of the 1D FET overlap in the linear regime with the dashed line of conductance $I_d/V_{ds} = G_Q = g_s g_v \cdot (q^2/h)$, which is the quantum of conductance.** The transition from this quantized-conductance regime to the saturation current regime is more abrupt in a 1D ballistic FET than in its 2D counterpart. This behavior was anticipated early on, in Chapter 5 for 1D transport, and can be seen explicitly from the last line in Equation 19.10 for $\eta_s >> +1$. Approximations similar to the 2D ballistic FET in the on- and off-states may be obtained from Equation 19.10 for the 1D ballistic FET:

$$I_{1d}^{lin} \approx [\frac{q^2}{h}\cdot g_s g_v]\cdot V_{ds} \text{ for } V_{ds} < V_{gs} - V_T$$

$$I_{1d}^{sat} \approx \frac{g_s g_v h[C_g(V_{gs}-V_T)]^2}{32qm_c^\star} \text{ for } V_{ds} > V_{gs} - V_T. \quad (19.11)$$

Unlike the DOS of 3D and 2D parabolic bands, the 1D DOS *decreases* with increase in energy as $1/\sqrt{E-E_c}$, and has a van Hove singularity at the band edge. As the gate voltage is increased, the barrier capacitance $C_b = \epsilon_b/t_b$ does not change, but the semiconductor quasi-Fermi

[22] The first boxed relation in Equation 19.10 must be self-consistently solved to obtain the 1D carrier density n_{1d} as a function of the gate voltage via $v_g = (V_{gs} - V_T)/V_{th}$, the drain voltage via $v_d = V_{ds}/V_{th}$, the barrier parameters via $n_b^{1d} = C_b \cdot V_{th}/q$, and the semiconductor band parameters via the 1D band-edge effective N_c^{1d}. Then, substituting $\eta_s = v_g - n_{1d}/n_b^{1d}$ from the first equation in the last boxed equation yields the complete output characteristics of the ballistic quantum wire FET.

level E_{Fs} is raised above E_c. Thus the semiconductor quantum capacitance $C_q \approx q^2 \cdot DOS$ *decreases*, and beyond a certain gate voltage dominates in the total gate capacitance given by $\frac{1}{C_g} = \frac{1}{C_b} + \frac{1}{C_q}$. This quantum capacitance-dominated regime of on-state operation increases the gate coupling to the quantum wire channel since then $C_g \approx C_q$. In 2D, the ballistic saturation current went as $I_d/W \sim (C_g V_{gs})^{3/2}$, but for 1D it goes as $I_d \sim (C_g V_{gs})^2$. In the regime in which quantum capacitance dominates, $C_g \approx C_q \sim 1/\sqrt{V_{gs}}$, which implies $I_d \sim (\sqrt{V_{gs}})^2 \propto V_{gs}$. This regime is of special interest in high-frequency amplifiers because the transconductance, or gain $g_m = \frac{\partial I_d}{\partial V_g}$ becomes *independent* of the gate voltage, enabling a highly linear operation regime without the generation of unwanted harmonics. Gain of FETs are discussed further in Section 19.6. Unlike in the 2D case, the velocity of mobile states in a single mode 1D channel FET is directed solely along the source-drain axis. This directionality is responsible for more efficient transport and efficiency of current flow per channel charge, since they are not moving at angles to the source-drain axis, as a significant fraction of carriers do in the 2D case. The better transport and gate control is offset by higher source and drain contact resistances to the single, or very few 1D modes of the channel.

19.5 The drift-diffusion FET

In the preceding sections of this chapter, we discussed the ballistic transistor as a limiting case of the FET when there is no scattering in the semiconductor channel. In this section we discuss the other limiting case, when the channel is much longer than the carrier mean free path between successive scattering events. Frequent scattering brings carriers into local equilibrium amongst themselves, rather than being in equilibrium with the source and drain contacts as was the case in the ballistic case. Fig. 19.14 shows a FET of channel length L, the corresponding calculated energy band diagrams, the mobile surface sheet charge $n_s(x)$ in the semiconductor, and the x-directed electric field in the channel. Because the carrier transport is scattering limited, they undergo drift and diffusion as introduced in Chapter 16, section 16.7, characterized by a mobility μ. We call it the "drift-diffusion" FET.

Let us consider an n-channel drift-diffusion FET. The local equilibrium of the mobile electrons at a point x from the source is characterized by a quasi-Fermi level $F_n(x) = -qV(x)$, where $V(x)$ is the local potential. Writing the electrostatic equation to relate the carrier density $n_s(x)$ at the point x in the channel to the gate voltage V_{gs} and the local electric potential $V(x)$, we obtain the gate-control equation[23]

$$\boxed{e^{\frac{n_s(x)}{n_b}}\left(e^{\frac{n_s(x)}{n_q}} - 1\right) = e^{\frac{V_{gs}-V_T-V(x)}{V_{th}}}.} \tag{19.12}$$

Because of the bias conditions, $V(x=0) = 0$ V at the source, and $V(x=L) = V_{ds}$ at the drain. The current density at any point x is

[23] Note that this equation is similar to the version used for the ballistic 2D FET earlier in Equation 19.2, with identical definitions of $n_b = C_b V_{th}/q$ and $n_q = C_q V_{th}/q$, and $V_T = \phi_b - \Delta E_c/q$ related to the M-O barrier height and the O-S conduction band offset. Since the scattering lets us define a local potential $V(x)$ at the semiconductor surface, it appears in the exponent and allows the determination of the local sheet density $n_s(x)$ for a given V_{gs}. In the drift-diffusion FET as well as the ballistic FET, $V(x) = 0$ at the source injection point when the source is grounded.

found from the local electric field $F(x) = -dV(x)/dx$:

$$\frac{I_d(x)}{W} = q\mu n_s(x)F(x) = q\mu n_s(x)[-\frac{dV(x)}{dx}]. \qquad (19.13)$$

To relate the local field to the local carrier density, we differentiate Equation 19.12 to get

$$dV(x) = -V_{th}\cdot\left[\frac{(\frac{1}{n_b}+\frac{1}{n_q})e^{\frac{n_s(x)}{n_q}}-\frac{1}{n_b}}{e^{\frac{n_s(x)}{n_q}}-1}\right]\cdot dn_s(x). \qquad (19.14)$$

Since the current is continuous at all x, $I_d(x) = I_d$, we use $dV(x)$ from Equation 19.14 in Equation 19.13 and rearrange to obtain

$$\int_{x=0}^{x=L} dx\cdot\frac{I_d(x)}{W} = -q\mu\int_{x=0}^{x=L} dV(x)\cdot n_s(x)$$

$$I_d = q\mu\frac{W}{L}V_{th}\int_{x=0}^{x=L} dn_s(x)\cdot\left[\frac{(\frac{1}{n_b}+\frac{1}{n_q})e^{\frac{n_s(x)}{n_q}}-\frac{1}{n_b}}{e^{\frac{n_s(x)}{n_q}}-1}\right]\cdot n_s(x)$$

$$I_d \underbrace{=}_{u(x)=n_s(x)/n_q} q\mu\frac{W}{L}V_{th}n_q^2\int_{u(x=0)}^{u(x=L)}\underbrace{\left[\frac{(\frac{1}{n_b}+\frac{1}{n_q})e^u-\frac{1}{n_b}}{e^u-1}\cdot u\right]}_{\frac{u}{n_b}+\frac{1}{n_q}\frac{ue^u}{e^u-1}}du$$

$$I_d = q\mu\frac{W}{L}V_{th}n_q^2\int_{u(x=0)}^{u(x=L)}\left[\frac{u}{n_b}+\frac{1}{n_q}\underbrace{\frac{ue^u}{e^u-1}}_{\approx u+e^{-u}}\right]du \qquad (19.15)$$

Note that we have defined $u(x) = n_s(x)/n_q$, which takes the values $u_s = n_s(x=0)/n_q$ at the source, and $u_d = n_s(x=L)/n_q$ at the drain, obtained by finding $n_s(x=0)$ and $n_s(x=L)$ directly from Equation 19.12, using the boundary conditions $V(0) = 0$ and $V(L) = V_{ds}$. The approximation to the last term is an excellent one for all u. Thus,

$$I_d = q\mu\frac{W}{L}(V_{th})n_q^2\int_{u_s}^{u_d}\left[(\frac{1}{n_b}+\frac{1}{n_q})u+\frac{1}{n_q}e^{-u}\right]du$$

$$\boxed{I_d = q\mu\frac{W}{L}V_{th}n_q\left[(1+\frac{n_q}{n_b})(\frac{u_s^2-u_d^2}{2})+(e^{-u_s}-e^{-u_d})\right]} \qquad (19.16)$$

is the desired drain current of the drift-diffusion FET as a function of the applied drain and gate biases and the device geometry. Fig. 19.15 shows the current transfer characteristics, and the output characteristics for several cases using Equation 19.16. Let us discuss the FET charge, field, and energy band diagrams in Fig. 19.14 and relate them to the transistor output characteristics in Fig. 19.15.

The first energy band diagram shown for $V_{gs} < V_T$ indicates E_F is inside the bandgap, with few carriers in the conduction band. The carrier density is significantly enhanced when $V_{gs} - V_T = +0.4$ V, when

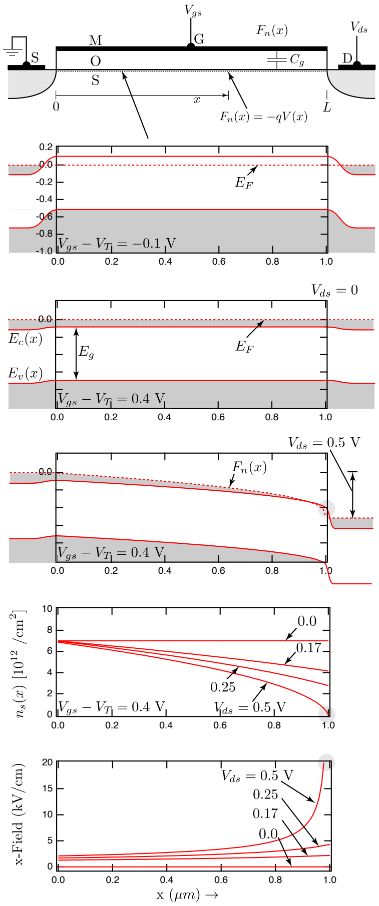


Fig. 19.14 Long channel FET energy band diagrams, mobile charge, and x-directed electric field distribution. The semiconductor bandgap is $E_g = 0.6$ eV, conduction band edge effective mass is $m_c^\star = 0.2m_e$ with a valley degeneracy $g_v = 1$, with electron mobility $\mu = 200$ cm^2/V·s. The source and drain are heavily doped n^{++}, and the oxide barrier is of thickness $t_b = 5$ nm and dielectric constant $\epsilon_b = 20\epsilon_0$. Note that when $V_{ds} = 0.5$ V $> V_{gs} - V_T = 0.4$ V, the carrier density at the drain end is "pinched off", and the x-directed electric field $F_x(x)$ rises sharply, all the while maintaining $I_d/W = qn_s(x)\mu F_x(x)$ constant from source to drain. The electron quasi-Fermi level in the drain end enters the bandgap in this region of high field, shown in gray circles.

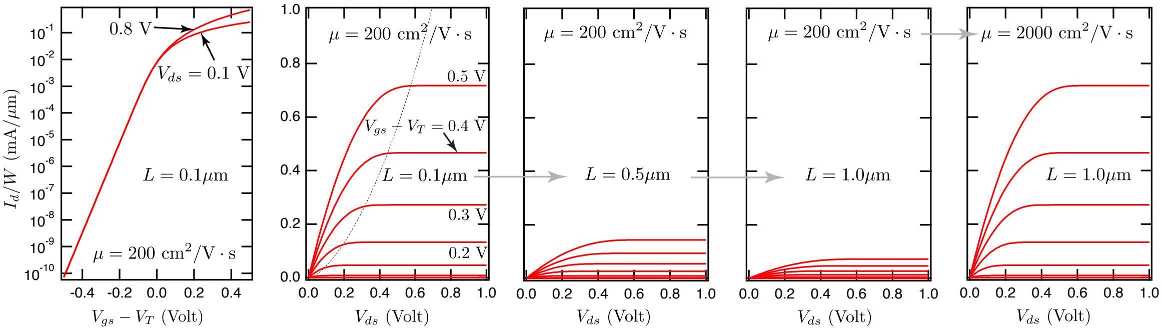


Fig. 19.15 Long channel FET characteristics. The gate dielectric has $\epsilon_b = 20\epsilon_0$ and is $t_b = 5$ nm thick. The semiconductor bandstructure information enters the gate control via an effective mass of $m_c^\star = 0.2m_e$ and a valley degeneracy of $g_v = 1$, and a lumped transport parameter of the mobility μ. The switching characteristics show the typical logarithmic dependence of the drain current on the gate voltage. For an electron mobility of $\mu = 200$ cm^2/V·s, the current drops from $I_d/W \sim 0.7$ mA/μm to ~ 0.07 mA/μm, as the gate length increased from $L = 0.1$ μm to 1.0 μm. This drop in the current drive may be recovered by increasing the mobility from 200 cm^2/V·s to 2000 cm^2/V·s. Note that velocity saturation, quite likely for gate length of $L = 0.1\mu$m, is neglected in this example for simplicity.

the Fermi level enters the conduction band. When a drain voltage V_{ds} is applied, the quasi Fermi level $F_n(x) = -qV(x)$ must change such that it drops by qV_{ds} from the source to the drain. Note that the potential difference across the oxide layer at x varies from the source to the drain: it is high at the source end, but for positive drain voltages, small on the drain side. Since the mobile carrier concentration $n_s(x)$ is proportional to the potential difference $V_{gs} - V_T - V(x)$ across the oxide capacitor, the carrier concentration must decrease from the source to the drain[24]. Because the current is constant in x, the x-directed electric field $F(x) = I_d/[q\mu n_s(x)]$ must increase from the source to the drain. Current saturation occurs for $V_{ds} > V_{gs} - V_T$ when the drain side of the channel is pinched off: this behavior is seen in Fig. 19.15 in the second panel, indicated by the dashed line[25]. Just as for the ballistic FET, the analytical approximations of the off-state subthreshold and on-state saturation current regimes of Equation 19.16 are obtained by using the two limits of Equation 19.3 of the carrier density $n_s(x)$:

$$\boxed{I_d^{off} \approx q\mu \frac{W}{L} n_q V_{th} e^{\frac{V_{gs}-V_T}{V_{th}}}} \text{ for } V_{gs} < V_T$$

$$\boxed{I_d^{on} \approx \mu \frac{W}{2L} C_g (V_{gs} - V_T)^2} \text{ for } V_{gs} > V_T \text{ and } V_{ds} > V_{gs} - V_T. \quad (19.17)$$

I_d^{off} goes exponentially with V_{gs}, and saturated I_d^{on} goes as its square.

19.6 CMOS and HEMTs

CMOS: The silicon MOSFET is by far the most widely used transistor. The heart of the device is the MOS capacitor, whose operation

[24] The calculated $n_s(x)$ is shown in Fig. 19.14 for various drain voltages: for the case chosen, the density decreases from the $V(x) = 0$ value of $\sim 7 \times 10^{12}$/cm^2 at the source to a very small value at the drain end at $V_{ds} = 0.5$ V. To obtain the carrier density profile $n_s(x)$, we use the continuity of the drain current yet again to write $I_d(x) = I_d(x = L)$ in Equation 19.16, and solve for $n_s(x)$ on the right by replacing $L \to x$. Using the profile of $n_s(x)$ obtained, the field is obtained from Equation 19.13 and the potential $V(x)$ is obtained from Equation 19.12. This directly yields the conduction band profile since $E_c(x) = qV(x) - k_b T \ln[e^{\frac{n_s(x)}{n_q}} - 1]$.

[25] The rest of the panels in Fig. 19.15 indicate the role of gate length scaling: because $I_d \propto \mu \cdot \frac{W}{L}$, increasing the gate length decreases the current drive. For some applications, it is necessary to increase L: for example in high-voltage transistors, since the maximum electric field in the channel must be kept below the critical breakdown field on the drain side. The last panel shows that increasing the mobility can help recover the reduction in current for longer gate lengths. This feature is obtained for example in high-voltage AlGaN/GaN HEMTs, where the critical field and electron mobility is higher than silicon.

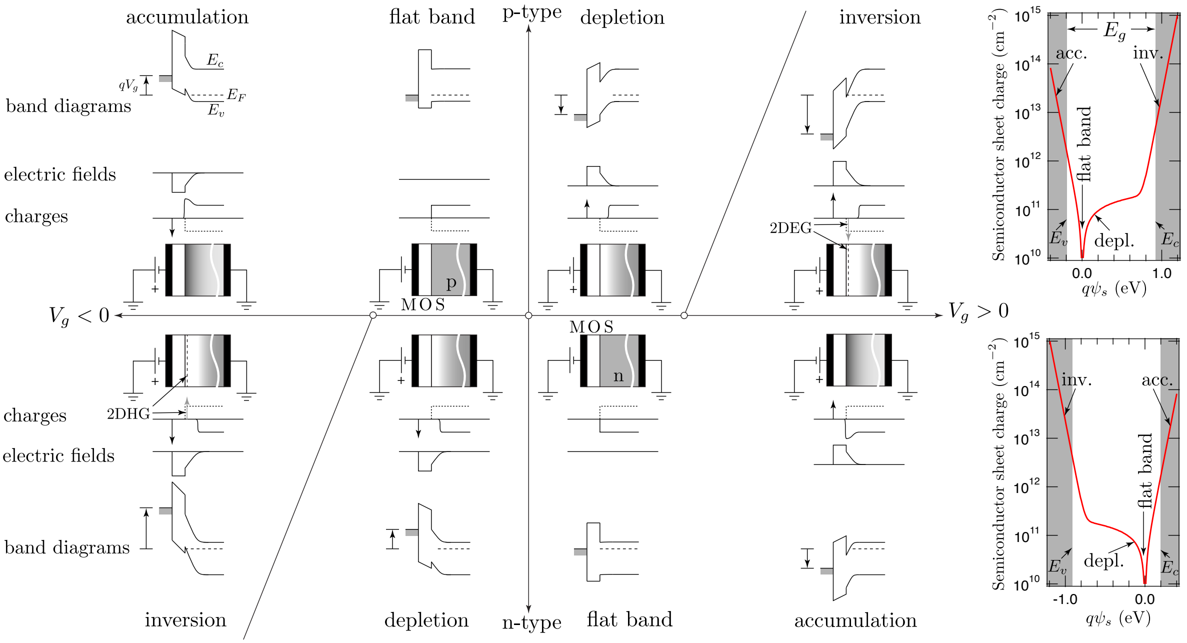

accumulation
flat band
p-type
depletion
inversion
band diagrams
E_c
qV_g
E_F
E_v
electric fields
charges
2DEG
p
MOS
$V_g < 0$
$V_g > 0$
MOS
n
2DHG
charges
electric fields
band diagrams
inversion
depletion
n-type
flat band
accumulation
Semiconductor sheet charge (cm^{-2})
10^{15}
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
E_g
acc.
inv.
flat band
E_v
depl.
E_c
0.0
1.0
$q\psi_s$ (eV)
inv.
acc.
flat band
depl.
-1.0
0.0

Fig. 19.16 MOS capacitor regimes of operation and semiconductor charge vs. surface potential for doping of $4 \cdot 10^{15}/\text{cm}^3$ at $T = 300\text{K}$.

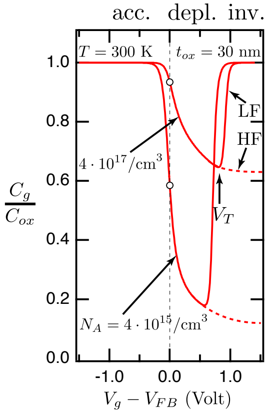


Fig. 19.17 The gate capacitance C_g of a nMOS capacitor potted as a fraction of the oxide capacitance. The low-frequency capacitance is obtained directly from $\frac{1}{C_g} = \frac{d\psi_s}{dQ_s} + \frac{1}{C_{ox}}$, which is derived in Equation 19.18 and the following text. Thus, $\frac{C_g}{C_{ox}} = \frac{1}{1+\frac{d\phi_s}{dQ_s}}$, and Fig. 19.16 shows Q_s vs. ψ_s.

regimes are shown in Fig. 19.16. Instead of the 2D semiconductor channel discussed in earlier sections, the bulk of the silicon (called the **body**) is doped p-type (upper half of Fig. 19.16), or n-type (lower half). Let us consider the p-type body MOS capacitor: this will lead to the n-channel MOS, or nMOSFET whose conducting channel is a 2D-electron gas that forms as an **inversion layer**: the semiconductor surface is converted from p-type in the bulk to n-type on the surface (i.e., inverted). The lower half of Fig. 19.16 shows the opposite, or the gate capacitor of the pMOSFET. The nMOS and pMOS together form the complementary MOS, or CMOS that powers much of digital logic and computation today.

Let us track the integrated sheet charge in the MOS capacitor for the nMOS (p-type body) shown in the right side plot of Fig. 19.16 by following the charge-field-band diagrams from the left to right. The corresponding measured gate capacitance shown in Fig. 19.17 is a powerful tool in the design and characterization of MOSFETs. When there are no net charges in the semiconductor or metal, there are no electric fields, and the energy bands are flat. This **flatband** condition is reached when the gate voltage is equal to the difference of the metal and semiconductor work function potentials, $V_{FB} = \phi_M - \phi_S$. This is shown in the upper half second column from the left in Fig. 19.16, and corresponds to zero net semiconductor charge in the plot on the right. When the net charge in the semiconductor is zero, the total band bending, or the **surface potential** of the semiconductor ψ_s is also zero.

When the gate voltage is made more negative, the negative charges in the metal plate of the MOS capacitor attracts more holes to the surface than in the flat-band equilibrium condition, causing a surface hole **accumulation layer** to form[26]. Moving to positive V_g, putting positive charge on the metal requires negative charge in the semiconductor for charge neutrality and to terminate the electric field lines originating from the positive charges in the metal. A p-type semiconductor can provide negative charges in two ways: first by **depletion**, exposing the negatively charged immobile ionized acceptors of density N_A^-, and if the surface potential is too large, by **inversion**, when mobile electrons form in the conduction band. To relate the charge in the semiconductor Q_s to the surface potential ψ_s under all of the above conditions: accumulation, depletion, and inversion, we solve Poisson's equation for a downward band bending $\psi(x)$[27]:

$$\frac{d^2\psi}{dx^2} = -\frac{q}{\epsilon_s}[p(x) + N_D - n(x) - N_A]$$

$$\frac{d^2\psi}{dx^2} = -\frac{q}{\epsilon_s}[p_{p0}e^{-\frac{\psi}{V_{th}}} - p_{p0} - n_{p0}e^{+\frac{\psi}{V_{th}}} + n_{p0}]$$

$$\underbrace{2\frac{d\psi}{dx}\cdot\frac{d^2\psi}{dx^2}}_{\frac{d}{dx}[(\frac{d\psi}{dx})^2]} = 2\frac{d\psi}{dx}\cdot(-\frac{q}{\epsilon_s})[p_{p0}(e^{-\frac{\psi}{V_{th}}} - 1) - n_{p0}(e^{+\frac{\psi}{V_{th}}} - 1)]$$

$$(\frac{d\psi}{dx})^2\Big|_{x_1}^{x_2} = -\frac{2qp_{p0}}{\epsilon_s}\int_{\psi(x_1)}^{\psi(x_2)} d\psi[(e^{-\frac{\psi}{V_{th}}} - 1) - \frac{n_{p0}}{p_{p0}}(e^{+\frac{\psi}{V_{th}}} - 1)]$$

$$\boxed{\frac{Q_s}{\epsilon_s} = F_s = \frac{\sqrt{2}V_{th}}{L_D}\sqrt{(e^{-\frac{\psi_s}{V_{th}}} + \frac{\psi_s}{V_{th}} - 1) + \frac{n_{p0}}{p_{p0}}(e^{\frac{\psi_s}{V_{th}}} - \frac{\psi_s}{V_{th}} - 1)}.} \quad (19.18)$$

The boxed equation uses Gauss's law $Q_s = \epsilon_s F_s$, where F_s is the electric field at the surface of the semiconductor, $L_D = \sqrt{\epsilon_s k_b T/q^2 p_{p0}}$ is the Debye length in the semiconductor, and ψ_s is the surface potential. It provides the relation between net semiconductor charge Q_s and ψ_s, which is plotted on the right side of Fig. 19.16. Since the excess gate voltage is related to the surface potential and the semiconductor charge by $V_g - V_{FB} = \psi_s + \frac{Q_s}{C_{ox}}$, we get the gate capacitance $C_g = dQ_s/dV_g$ by simply differentiating this equation w.r.t. Q_s to obtain $\frac{1}{C_g} = \frac{d\psi_s}{dQ_s} + \frac{1}{C_{ox}}$, which is plotted in Fig. 19.17. The gate capacitance approaches C_{ox} for both accumulation and inversion, and drops in the depletion regime. The threshold voltage for the inversion channel to form is higher for larger N_A. The gate capacitance is different in the low- and high-frequency regimes[28].

Fig. 19.18 shows how the conductive inversion layer at the surface of the MOS capacitor is contacted by the appropriate heavily doped source/drain regions to create the MOSFET. Note that the S/D contact regions are **doped the opposite type to the semiconductor body** to

---

[26] The excess accumulation hole sheet density rises exponentially with the surface potential ψ_s, which is now negative. The electric field points from the positively charged holes to the negative charge on the metal, with most of the voltage drop occurring in the oxide layer. The corresponding measured gate capacitance at accumulation is therefore expected to be, and indeed is, close to the parallel-plate value $C_{ox} = \epsilon_{ox}/t_{ox}$ as seen in Fig. 19.17 which we derive quantitatively below.

[27] Here we have used that deep inside the semiconductor bulk, $N_D - N_A = n_{p0} - p_{p0}$, $p_{p0} = N_A$ and $p_{p0}n_{p0} = n_i^2$. For all x, $p(x) = p_{p0}e^{-\frac{\psi}{V_{th}}}$ and $n(x) = n_{p0}e^{+\frac{\psi}{V_{th}}}$. We substitute these in the Poisson equation to solve it. To evaluate the integral, we use $x_1 = +\infty$ is deep inside the semiconductor bulk, and $x_2 = 0$ is the surface of the semiconductor. Since $\psi(x)$ is the *downward band bending*, it controls the electron density as $n(x) = n_{p0}e^{+\psi(x)/V_{th}}$ and thus ψ_s is the net surface band bending, or the surface potential. Because electrons pile up as ψ becomes more positive, the surface of the semiconductor is inverted as ψ_s approaches the bandgap of the semiconductor.

[28] The surface inversion layer of electrons in the p-type semiconductor must form by thermal generation, which takes time. As shown in Fig. 19.17, the inversion capacitance is different if the gate votage is swept at a low frequency (LF), which gives the inversion layer time to form. In this case, $C_g \to C_{ox}$. When the gate voltage is swept at a high frequency (HF) condition, the inversion layer does not form. This property enables the use of MOS capacitors in dynamic random access memories (DRAM). See Exercise 19.9 for more.

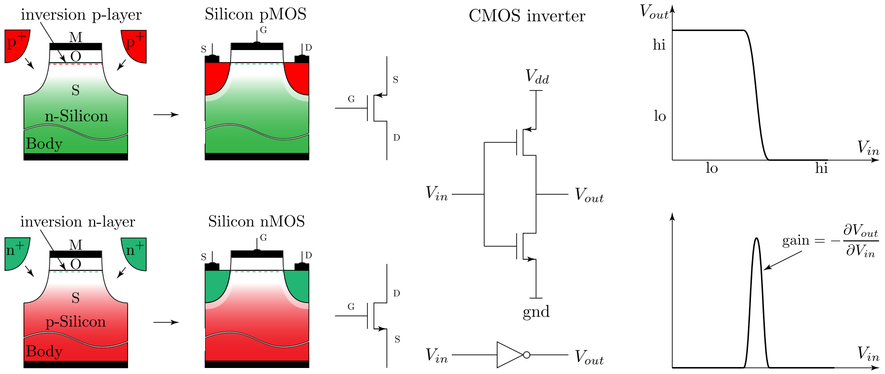


Fig. 19.18 From MOS capacitors to pMOS and nMOS FETs to the CMOS inverter.

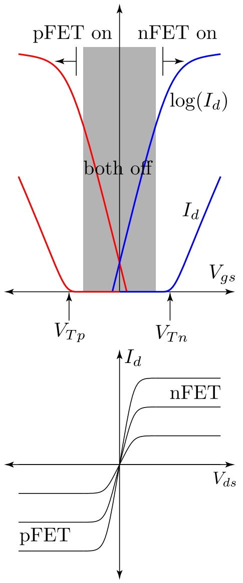


Fig. 19.19 Operation regimes of a nFET and a pFET. The transfer characteristics on top show the drain currents in linear and log scales as a function of the gate voltage. The nFET turns on for $V_{gs} > V_{Tn}$ and the pFET for $V_{gs} < V_{Tp}$. The output characteristics of I_d vs. V_{ds} are shown at the bottom. Fig. 19.18 shows the CMOS configuration using the nMOSFET and pMOSFET.

form a low-resistance contact to the inversion layer. The inversion layer is only present when $V_{gs} > V_{Tn}$ for the nMOSFET, and $V_{gs} < V_{Tp}$ for the pMOSFET. Under these conditions, current flows through the conducting inversion layer between the source and the drain contacts as indicated Fig. 19.19.

The nMOS and pMOS FETs connected as shown in Fig. 19.18 form the complementary MOS, or **CMOS inverter**. The gates of the pMOSFET and nMOSFET are shorted to form the input terminal, and the drains are shorted to form the output terminal of the inverter. The source contact of the pMOSFET is connected to a positive constant drain voltage V_{dd} line, and the source of the nMOSFET is connected to the ground line. When the input voltage on the gate is low ($V_{in} = 0$), the nMOSFET is off since $V_{gs} = 0 < V_{Tn}$ as seen in Fig. 19.19. However, since for the pMOSFET $V_{gs} = V_{in} - V_{dd} < V_{Tp}$, the pMOSFET is on, and the pMOS channel resistance is low. The pMOSFET therefore brings the potential V_{out} of the output terminal to V_{dd}, which is a high voltage. On the other hand, when V_{in} is high, the pMOSFET turns off, but the nMOSFET turns on, pulling the potential V_{out} to its source potential, or ground. Thus, the circuit converts a low V_{in} or a logical **zero** to a high V_{out} or a logical **one**, and a logical one to a logical zero, earning this circuit the name of a logical **inverter**. The input-output transfer curve is shown in Fig. 19.18, and its derivative $-\frac{dV_{out}}{dV_{in}}$ in the transition regime defines the gain of the CMOS inverter. The CMOS inverter is a very energy efficient circuit, since there is power dissipation only when the input bit flips; very little power is dissipated to hold the output bit as the inverted input bit when there are no bit flips at the input.

Now the MOSFET device can be a long-channel FET, or a very short channel device which is nearly ballistic. Fig. 19.20 shows the energy band diagram under the gate, another from the source to the drain, and also a k-space occupation function at the source injection point as discussed earlier in this chapter. The major difference from a 2D

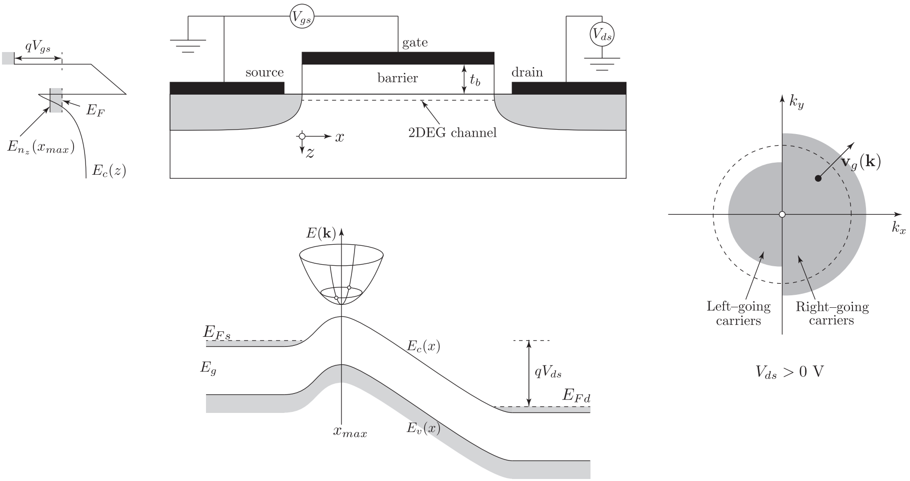

qV_{gs}
E_F
$E_{n_z}(x_{max})$
$E_c(z)$
V_{gs}
gate
source
barrier
t_b
drain
V_{ds}
2DEG channel
x
z
$E(\mathbf{k})$
E_{Fs}
E_g
$E_c(x)$
qV_{ds}
E_{Fd}
x_{max}
$E_v(x)$
k_y
$\mathbf{v}_g(\mathbf{k})$
k_x
Left–going carriers
Right–going carriers
$V_{ds} > 0$ V

Fig. 19.20 Illustrating the physics of a ballistic FET under nonzero drain bias. Field effect transistor, energy band diagram, and **k**-space occupation of states.

semiconductor channel is the presence of the depletion thickness barrier and potentially several subbands in the triangular quantum well at the semiconductor surface, the lowest of which is labeled as $E_{n_z}(x_{max})$. The model discussed in the earlier sections of this chapter apply to the MOSFET both in the ballistic limit, and the drift-diffusion limit because the strong confinement of the carriers justifies treating them as 2D systems.

Fig. 19.21 shows the practical structure of modern planar MOSFETs, and its vertical embodiment, the FinFET. The gate stack consists of a high-κ/metal gate with a spacer covering the sidewalls. In the planar MOSFET, the electric field is exerted by the gate charges vertically into the semiconductor surface. The current flows between the source and the drain on the surface inversion sheet of width W under the gate of length L, defining the W/L that appears in the drift-diffusion FET model. Because of the need for better electrostatic control of the conducting channel for very small gate lengths, new MOSFETs use channels shaped as a fin, which turns the channel by 90°. The gate exerts the field through the sidewalls of the fins. For this reason, this geometry is referred to as the tri-gate MOSFET, or FinFET. An increase in transistor density can be obtained by bringing the fins closer to each other. Furthermore, the fin may be broken up into several nanowires (or 1D quantum wires), and gates can be wrapped around each wire, forming stacked gate-all-around (or GAA) FETs.

HEMTs: Fig. 19.22 shows the cross-section construction of high electron mobility transistors (HEMTs), which are also called heterostructure field-effect transistors (HFETs). HEMTs typically use III-V semiconductors with much higher electron mobilities. The electrons are typically located in a quantum well at an epitaxial heterojunction which is completely crystalline, as opposed to the silicon MOSFET, where the

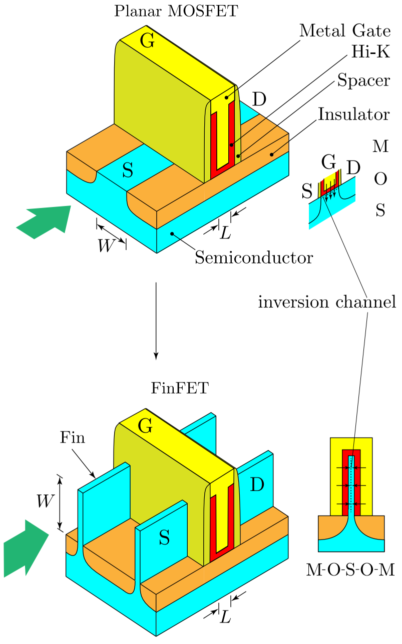

Planar MOSFET
G
Metal Gate
Hi-K
Spacer
D
Insulator
M
O
S
S
G
D
S
W
L
Semiconductor
inversion channel
FinFET
Fin
G
D
W
S
L
M-O-S-O-M

Fig. 19.21 Planar MOSFET and FinFETs. The conducting channel in FinFETs is inside the fin, and gated from the sides. The FinFET geometry was introduced by Hisamoto, Kaga, and Takeda in 1991 (see Exercise 19.2).

inversion channel is formed at the interface between the crystalline semiconductor and the amorphous oxide. Because of this reason, the electrons in the HEMT suffer far lower interface roughness scattering. Furthermore, the mobile electrons in HEMTs are spatially separated from dopants, reducing the Coulomb scattering. The combined reduction of scattering mechanisms leads to typically high mobilities in HEMTs. The mobile carriers are provided for example by modulation doping as shown for InAlAs/InGaAs HEMTs, or by polarization-induced doping in AlGaN/GaN HEMTs. Because of the high mobilities and high electron saturation velocities, for the same gate length as MOSFETs, HEMTs typically are faster. Compared to silicon nMOS, AlGaN/GaN HEMTs have 10 times higher mobilities, and at the same time offer much higher breakdown voltage capability. These properties make HEMTs attractive for high-speed transistor applications as high-frequency RF amplifiers, and high speed switching of large voltages for energy-efficient power electronics.

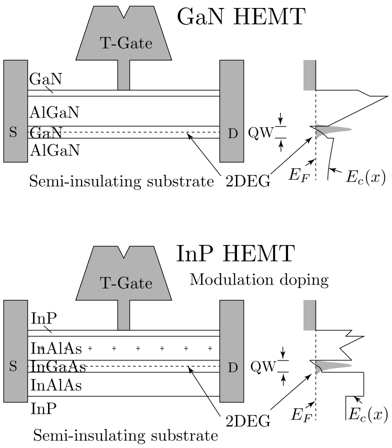


Fig. 19.22 Layer structures and energy band diagrams of GaN and InP HEMTs showing the quantum well (QW).

Performance metrics: Fig. 19.23 shows the performance metrics of a general FET. In addition to the highest possible drain current, the maximum drain voltage V_{br} that may be applied without breaking down the transistor is an important metric. Since this increases the net area under the $I_d - V_{ds}$ curve, it enables high power amplifiers in high-speed electronics, and also high-voltage switching for power electronics. While the maximum on-currents are typically in the range of 1 mA/μm, V_{br} can vary widely for different semiconductor technologies and device geometries; in general, a larger bandgap semiconductor leads to a higher V_{br}. The maximum power is $P_{max} \approx \frac{1}{8} \cdot I_d^{max} \cdot (V_{br} - V_{knee})$, a part of the shaded rectangle in Fig. 19.23. Fig. 19.23 shows the transfer curve I_d vs. V_{gs}, and its derivative, which is defined as the **transconductance** of the FET, $g_m = \frac{\partial I_d}{\partial V_{gs}}$. The transconductance takes the role of the gain of the FET; for example, if a load resistance R_L is connected to the drain, a change in the gate voltage ΔV_{gs} appears across this load as $(\Delta I_d)R_L$, implying the voltage gain of the FET is $\frac{(\Delta I_d)R_L}{\Delta V_{gs}} = g_m R_L$. The transconductance, like the current, is expressed per unit length; typical values for high-performance FETs are in the $0.1 - 1.0$ mS/μm range. Writing the saturation current as $I_d/W \sim qn_s v_{sat}$ where v_{sat} is an ensemble saturation velocity, we get $g_m \sim C_g v_{sat}$, since $C_g = d(qn_s)/dV_{gs}$.

Fig. 19.23 shows the voltage and power gain of a FET as a function of the frequency of the voltage signal applied at the gate terminal. The intrinsic speed of the FET is related to the time τ_{ch} it takes for the electron to traverse the channel of length L via $f_T = \frac{1}{2\pi\tau_{ch}} = \frac{v_{sat}}{2\pi L}$. The voltage gain $g_m R_L$ goes to unity at this characteristic frequency, called f_T. The maximum frequency at which power ($\propto I_d \cdot V_{ds}$) amplification can be obtained is called f_{max}. Due to the presence of additional delays due to capacitances and resistances extrinsic to the device, a lower f_T is typically obtained. For FETs, the voltage gain cutoff frequency f_T

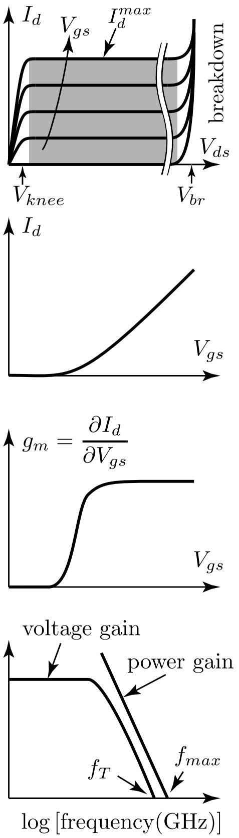


Fig. 19.23 Performance metrics of a FET.

and the unilateral power-gain cutoff frequency f_{max} are given by

$$f_T = \frac{1}{2\pi\tau} = \frac{1}{2\pi(\tau_{source} + \tau_{ch} + \tau_{drain})}$$

$$f_T = \frac{g_m}{2\pi[(C_{gs} + C_{gd})(1 + g_{ds}R_s) + C_{gd}g_mR_s]} \approx \frac{g_m}{2\pi C_{gs}} \approx \frac{v_{sat}}{2\pi L}$$

$$f_{max} = \frac{f_T}{2\sqrt{g_{ds}(R_g + R_s + r_{ch}) + g_mR_g\frac{C_{gd}}{C_{gs}+C_{gd}}}}. \quad (19.19)$$

The gate resistance[29] does not appear in f_T. As an example, for $v_{sat} \sim 10^7$ cm/s and $L = 50$ nm, the intrinsic $f_T \sim 320$ GHz, but due to parasitics it can typically be lowered to ~ 200 GHz. The value of the f_{max} is sensitive to the gate resistance, requiring special mushroom, or T-shaped gates as shown in Fig. 19.22 to keep the gate length small at the foot, but at the same time lower the gate resistance by making the head fat. Modern very short gate length CMOS and HEMTs can reach $f_T \sim 0.5 - 1.0$ THz and f_{max} in the same range. A primary impediment to even higher speeds are short-channel effects such as drain-induced barrier lowering (Exercise 19.11), due to which I_d fails to saturate, leading to large g_{ds}. The other impediment is that contact resistances degrade the gain (Exercise 19.15), and are notoriously difficult to lower beyond certain limits posed by quantum mechanics of crystalline solids, which we discuss next.

19.7 Source/drain ohmic contacts

We discussed the construction of tunneling ohmic contacts in Chapter 18, Fig. 18.11. The goal is to make the source resistance R_s and drain resistance R_d of the FET and the resulting voltage drop $I_d(R_s + R_d)$ in the contacts is as small as possible compared to the drain voltage V_{ds} that drives the current. In Chapter 17, Section 17.2 we discussed that there is a quantum limit of how low a resistance a semiconducting or metallic channel in a crystalline solid can attain. The maximum conductance for a 1D mode is $2q^2/h$, and conductances of several modes add to give a net contact resistance normalized to the width of a 2D conducting channel $R_c \cdot W \sim \frac{h}{2q^2k_F} \sim \frac{0.026}{\sqrt{n_{2d}}}$ k$\Omega \cdot \mu$m. Experimental contact resistances can today reach close to the quantized conductance limits, as indicated in Fig. 19.24. Achieving the low contact resistance limits requires significant efforts in doping and metal choices catered towards the chemistry and physics of each semiconductor/metal choice. This quantum limit constrains the speed and performance on the latest generation of nanoscale FETs, both in digital and high-speed amplifier applications. We saw in the earlier sections of this chapter that the best nanoscale FET channels are capable of carrying currents in the $I_d \sim 5$ mA/μm range at $V_{ds} \sim 0.5$ V in the ballistic limit. The voltage drop at the source and drain contacts in the quantum limit for $n_{2d} = 10^{13}/\text{cm}^2$ is then $V_c = \frac{I_d}{W}(R_s \cdot W + R_d \cdot W) =$

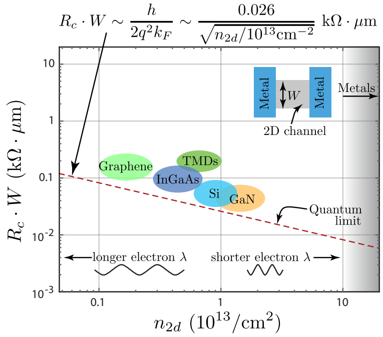


Fig. 19.24 Source and drain contact resistances and their quantum limits for crystalline semiconductor FETs with 2D conducting channels. Since each 1D mode can have a conductance maximum of $2q^2/h$, the resistance of a 2D channel composed of many 1D modes is $R_c \cdot W \sim \frac{h}{2q^2k_F} \sim \frac{0.026}{\sqrt{n_{2d}}}$ k$\Omega \cdot \mu$m scaled to the width of the contact. The resistance is between the heavily doped 3D regions and the 2DEG like in the MOSFET.

---

[29] The gate-source capacitance C_{gs} is desired, and the primary extrinsic undesired capacitance is the gate-drain capacitance C_{gd}. The undesired resistances are the source resistance R_s, the resistance of the gate metal R_g, and the output conductance g_{ds}. The change in the drain current cannot follow the gate voltage change any faster than it takes the electron to go from the source to the drain. This offers an intuitive explanation for $f_T \approx v_{sat}/2\pi L$ in Equation 19.19.

$I_d \cdot 2R_c \sim 0.26$ V, which is unfortunately a **significant fraction of** V_{ds}. Even at a current of 1 mA/μm, the voltage drop at the contacts is far from a negligible fraction of V_{ds}. This is why a significant effort in lowering the contact resistances is needed to harvest the fruits of scaling the channel lengths, because $I_d^2(R_s + R_d)$ is an irrecoverable loss of energy in the device. This loss increases fast for several billion transistors!

The source and the drain contacts are designed for as low resistance metal-semiconductor ohmics as possible, by making use of heavy doping in the semiconductor and choosing low-barrier height metals. Because in an n-type channel FET the current is carried by mobile electrons in the conduction band, a low-resistance ohmic contact is necessary to the conduction band of the semiconductor. For a p-channel FET a low-resistance ohmic contact is necessary to mobile holes in the valence band. As a result, a typical choice for an nFET ohmic contact is a metal whose Fermi level E_{FM}^n nearly lines up with the conduction band edge E_c of the semiconductor, so that the Schottky barrier height is the smallest possible. The n-contact resistance is lowered further by doping the semiconductor with donors heavily near the metal/semiconductor junction to reduce the depletion region to a few nms, which enables efficient tunneling. Similarly, for a pFET the metal Fermi level E_{FM}^p must line up as close as possible to the valence band edge E_V of the semiconductor. The ideal n-ohmic metal and p-ohmic metal for a given semiconductor[30] must therefore have the difference of their work functions close to the bandgap of the semiconductor $E_{FM}^n - E_{FM}^p \approx E_c - E_v = E_g$. The practical choice of ohmic metals involves several factors concerning the interface, chemical, and processing stability and reliability. Though the contact resistance for all FETs should be made as small as possible, it is especially necessary for short-channel FETs. This is because as the distance from the source to the drain is scaled down, the channel resistance reduces and approaches the ballistic limit, but the contact resistance does not reduce. The contact resistance is not modulated by the gate either. This causes an all-around loss of control and energy efficiency of the transistor. As we discussed in the last section, the contact resistances significantly restrict the speed of switching and amplification by the FET. If the S/D distance is very long, and the resistance of the channel dominates over the contact resistance, the deleterious effects of the contact resistance are partially suppressed[31].

19.8 A brief history of FETs

The idea of the field-effect transistor was conceived much earlier than bipolar transistors. The earliest patent was filed in 1925 by Lilienfeld (Fig. 19.25), who had moved from the University of Leipzig in Germany to the United States for his entrepreneurial pursuits. The patent put forward the idea that if the conductivity of a material could be

[30] Because the work functions of elemental metals range from $2.5 - 6.0$ eV, it is difficult to find optimal choices for both n- and p-contacts to very wide bandgap semiconductors such as nitrides (e.g. AlGaN and AlN) and diamond whose bandgaps can exceed 5 eV.

[31] This is similar to a typical dilemma of choosing air travel over trains: most of the time spent in flying a short distance goes into getting to the airport and checking in, boarding, and getting off: the flight time is 'ballistic', or minuscule in comparison. This is analogous to a short channel, or ballistic FET: a significant portion of the delay and voltage drop is in the contacts, and not in the channel. For traveling a long distance, flight is an easy choice since the flight time is much longer than the waits at the airports, mimicking the fact that the channel resistance and delay is longer than in the contacts for a long-channel FET.

modulated by field-effect, it could perform the same operation as the electronic amplification that was at the time achieved by using vacuum tubes. The understanding of semiconductors and their underlying conductivity occured in the early 1930s, after Bloch's concepts of electron wave transport in crystals, and Alan Wilson's subsequent explanation of the difference between metals and semiconductors based on filled and empty bands. In 1945, Heinrich Welker (Fig. 19.26) patented the idea of a junction-FET or JFET. Research groups at the Bell laboratories attempted to demonstrate the field-effect phenomena driven by William Shockley's ideas, but the electrons in dangling bond surface states thwarted these efforts. The electric field lines that were supposed to penetrate into the semiconductor and contribute to the modulation of conductivity were essentially terminated at the surface states. That the surface states were responsible for the lack of field-effect was identified by John Bardeen, after examining several experimental efforts of his colleague Walter Brattain. In fact, they could demonstrate a weak field-effect with electrolyte gating, but not in the solid state. In these investigations, they chanced upon a different kind of device, the point contact transistor, the precursor to the first semiconductor bipolar junction transistor, discussed in Chapter 18. Motivated by the idea that the depletion region of a reverse-biased pn junction increases with the reverse-bias voltage, Shockley in 1952 quantitatively evaluated the output characteristics of a JFET and offered a blueprint of its design. A JFET was subsequently demonstrated by Dacey and Ross at the Bell Laboratories in 1953 as the first operational FET. Meanwhile, in 1953 when investigating the properties of n-p-n bipolar transistors with a floating collector, W. L. Brown in Bell Laboratories made the surprising observation that a floating collector terminal could attain the same potential as the emitter, in spite of being separated by a p-layer. He was able to prove that this was because the surface of the p-type layer was **inverted**, forming a conducting electron channel. Soon after, efforts to electrically passivate the surface states using SiO_2 succeeded, significantly lowering the surface recombination currents that hurt the performance of the rapidly advancing state-of-the-art bipolar junction transistors. In 1959, Mohammed Atalla and Dawon Kahng (Fig. 19.27) working at the Bell Laboratories were able to stabilize the surface of silicon and succeeded in realizing the first silicon MOSFET. In 1963, CT Sah and Frank Wanlass (Fig. 19.28) of Fairchild Semiconductors conceived of and demonstrated the first complementary MOS (CMOS) logic inverter. Several variants of the FET soon followed, for example Carver Mead demonstrated the metal-semiconductor FET (MESFET) using a Schottky gate in 1966, and Takashi Mimura introduced the modulation-doped high-electron mobility transistor (HEMT) in 1979. In 1980, a fundamentally new physical phenomenon, the integer quantum Hall effect was observed in FETs, followed by the fractional quantum Hall effect in 1982 (these are discussed in Chapter 25). To date, more than 10^{22} transistors have been fabricated, which is more than the number of stars in the Milky Way galaxy! It is likely the largest num-

Fig. 19.25 Julius Lilienfeld filed the earliest patent for the field-effect transistor in 1925, before the physics of semiconductors, or even the fundamentals of quantum mechanics were understood.

Fig. 19.26 Heinrich Welker, Arnold Sommerfeld's student, realized the earliest junction-Field Effect Transistor, or JFET. Welker was a pioneer in III-V semiconductors and heterostructures.

Atalla
Kahng

Fig. 19.27 Atalla and Kahng realized the first MOSFET in Bell Laboratories in 1959.

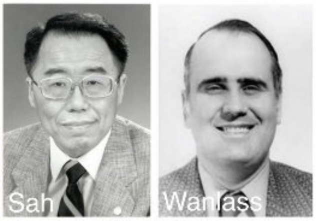


Fig. 19.28 In 1963, CT Sah and Frank Wanlass realized and demonstrated the remarkable properties of the CMOS circuit by combining the nMOS and pMOS in its now iconic format.

ber of devices that have been manufactured in the micro- and nanoscale to exacting standards by the human race. The FET has changed way logic, memory, and communication is performed on earth (and beyond!) in a profound manner.

In addition to fundamentally changing the fabric of logic, memory, and communication systems, the FET has contributed to significant discoveries in several other fields, such as in radio astronomy, weather prediction and metrology, medicine and biology, and imaging and sensing (e.g. the CCD camera is based on MOSFETs). The increased computing power by scaling of FETs enabled significant advances in the capacity to train computers to "think" and "learn" (so-called machine-learning), and is advancing artificial intelligence. In fact, in the past few decades the FET has continued to be "written off" as impossible to improve. But every few years there have been breakthroughs, and FETs have emerged better, faster, more efficient, and *far* more capable.

19.9 Chapter summary section

In this chapter, we learned:

- The FET operates by gate voltage control of the drain current by controlling the number of mobile carriers in a conductive channel by field-effect. The drain current is controled exponentially below a threshold voltage $I_d \sim e^{V_{gs}/V_{th}}$, and in a polynomial fashion above threshold, $I_d \sim (V_{gs} - V_T)^n$. In the on-state, the carrier density for 2D channel FETs is $\sim 10^{13}/\text{cm}^2$, and the on-state current densities are $\sim$ mA/μm, even when there is no scattering in a ballistic FET.
- If there is significant scattering in the channel, the carrier mobility plays an important role, and so does the gate length, since $I_d \sim \mu \frac{W}{L}$. Then, scaling the FET to smaller gate lengths improved the device performance.
- The silicon MOSFET operates with an inversion channel as the conducting layer. Since both nFETs and pFETs can be realized on the same wafer, it enables energy efficient complementary MOS, or CMOS circuits. On the other hand, high electron mobility transistors (HEMTs) based on III-V semiconductor heterostructures are used for power amplifiers that can operate at high efficiencies at very high frequencies.

Further reading

Sze's **Physics of Semiconductor Devices** treats the MOSFET and the zoo of various FETs most comprehensively using the drift-diffusion theory. Grove's **Physics and Technology of Semiconductor Devices** is a classic. **Fundamentals of Modern VLSI Devices** by Taur and Ning has a modern treatment, and includes in some chapters the ballistic FET models. The ballistic FET models are most extensively discussed in the short book **Nanoscale Transistors** by Lundstrom and Guo. I strongly recommend reading Natori's 1994 Journal of Applied Physics paper in which he introduced the ballistic FET model. Every device physicist has a personal favorite book for (mostly) unexplained reasons, mine is **Device Electronics for Integrated Circuits** by Muller, Kamins, and Chan.

Exercises

(19.1) **Practice problem: long channel FETs**
Equations 19.17 give the standard "long-channel" model current-voltage characteristics of FETs as a function of gate voltage. Answer the following questions:

(a) Calculate the saturated on-current for a channel mobility of $\mu = 150$ cm^2/V·s, $t_b = 10$ nm, $\epsilon_b = 3.9\epsilon_0$, $L = 1$ μm, and $W/L = 10$ at $V_{gs} - V_T = 1$ Volt.

(b) Note that the saturated current in part (a) does not depend on V_{ds}. For what values of V_{ds} is this approximation correct?

(c) Typical saturation current densities per unit width I_d^{on}/W approach 1 mA/μm in most FETs today. How does the above calculation compare?

(d) Calculate the off-state current per unit width I_d^{off}/W for the same dimensions as above, and $V_{gs} - V_T = -1.0$ V. What is the on/off ratio I_d^{on}/I_d^{off} for $V_{gs} - V_T = \pm 1$ Volt?

(e) Examine the impact of the geometric term W/L in the drive current of the FETs. How can one maximize the on-current for a given voltage, and what are the tradeoffs? The ration W/L is an important design parameter in the layout of transistors in circuits in microprocessors.

(19.2) **Shedding the fat: DELTA**
In 1991, Hisamoto, Kaga, and Takeda of Hitachi Central Laboratories in Japan introduced the DELTA, a fully **de**pleted **l**ean-channel **tra**nsistor. Fig. 19.29 shows the FET, an energy band diagram, and the sub-threshold slope as a function of the gate length. This work is one of the earliest demonstrations of the FinFET geometry of transistors. Answer the following questions about this work.

(a) Read the paper critically, and note down the advantages of this geometry as described by the authors to prevent several problems encountered in planar transistors.

(b) Discuss how they prove these advantages experimentally in the output characteristics of the FETs. Most of the advantages stem from removing the "bulk" semiconductor, which is the effective "fat" in the device.

(c) Describe how the design of today's FinFETs with channel lengths of below 10 nm can trace its origins back to this initial demonstration.

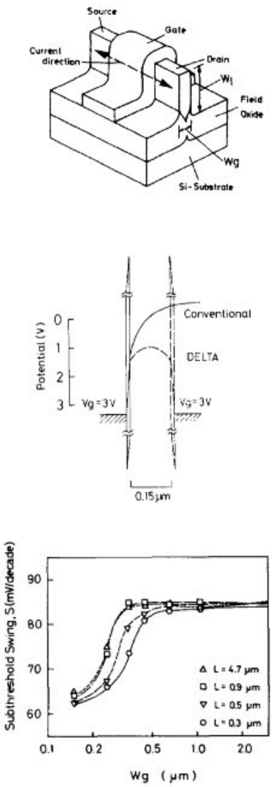


Fig. 19.29 The DELTA FET, from IEEE Transactions on Electron Devices, vol. 38, page 1399 (1991). This work is one of the earliest justifications and experimental demonstrations of the advantages of the FinFET.

(19.3) **The ballistic field-effect transistor**

In the chapter, we derived the characteristics of a ballistic field-effect transistor. You are going to fill in a few steps, and solve a closely related problem.

(a) Make log-scale and linear-scale plots of the gate-induced 2D electron gas (2DEG) carrier density at 300 K and 77 K vs. the gate voltage V_{gs} of FETs for an insulating barrier of $t_b = 2$ nm, and $\epsilon_b = 10\epsilon_0$ for three semiconductor channels: one that has $m_c^\star = 0.2m_0$, $g_s = 2$, and $g_v = 2$, the second has $m_c^\star = 0.2m_0$, $g_s = 2$, and $g_v = 1$, and the third has $m_c^\star = 0.05m_0$, $g_s = 2$, and $g_v = 1$. What is the difference? Compare with the figures in this chapter.

(b) Show why the ballistic current density is given by $J_{2d} = J_0[\mathcal{F}_{1/2}(\eta_s) - \mathcal{F}_{1/2}(\eta_s - v_d)]$, where $J_0 = q \cdot \frac{\hbar N_c}{m_c^\star}$, and all symbols have their usual meanings as they appear in the book, and N_c is the effective conduction band edge DOS.

(c) Make a plot of the ballistic FET currents of the three semiconductors of part (a). Make the drain current I_d vs. gate voltage V_{gs} in the linear and log scales, and the drain current I_d vs. drain voltage V_{ds} plots, similar to those shown in the figures in this chapter.

(d) Describe qualitatively what sorts of changes in the device characteristics would you expect if instead of the 2DEG channel, you had a 1D channel in the ballistic FET. Remember you have shown before that the ballistic conductance per 1D channel is limited to the quantum of conductance $G_0 = g_s g_v \frac{q^2}{h}$, where h is the Planck's constant.

(19.4) **Double-gate and gate-all around FETs**

In FETs made with semiconductor channels in the form of thin membranes such as in silicon-on-Insulator (SOI) or for 2D semiconductor crystals, gates above and below the plane of the semiconductor channel (see Figures 19.5 and 19.6) provide stronger gate control of the channel conductivity. Answer the following questions:

(a) Show that for symmetric barriers on top and the bottom with dielectric constant ϵ_b and thickness t_b, the gate capacitance C_g is twice of a single barrier of the same material and geometry.

(b) Show that the long-channel saturated current drive per unit width I_d^{on}/W increases for the same $V_{gs} - V_T$. Show that this boost in current drive with higher C_g remains true for the ballistic FET characteristics (Equations 19.7), though the quantitative dependence is different.

(c) The increased gate control is crucial in making FETs with shorter channel lengths and reduced dimensions. For 1D channels, wrapping the gate all around the channel (in the gate-all-around geometry) allows scaling to short channel lengths. Discuss the progress made in this direction in nanoscale FETs today, and the challenges an opportunities in their realization.

(19.5) **Barrier and quantum capacitance ratios**

In the quantum mechanical ballistic model of the FET, the on-current depends on the ratio of the carrier and quantum capacitance C_b/C_q in a non-monotonic manner. Examine this dependence, and as discussed in the text, show that for an optimal ratio the drive current is maximized. Discuss how this optimal choice dictates how the design of the barrier dielectric material and its thickness is tied to the electron bandstructure properties of the semiconductor channel.

(19.6) **Jump-start: the injection velocity**
Equation 19.8 showed that the injection velocity v_{inj} at the source side of the channel in a ballistic FET was found to play a central role in its output characteristics.

(a) Discuss what properties of the semiconductor at the source are desirable to maximize the saturated injection velocity v_{inj}^{sat} for a given $V_{gs} - V_T$. Note that the gate capacitance C_g also includes the properties of the semiconductor channel via the quantum capacitance.

(b) Narrower bandgap semiconductors with small $m_c^\star$ boost the injection velocity. But what limitations do they pose in achieving other desirable transistor characteristics (e.g. a high on/off ratio)?

(19.7) **Linearity of field-effect transistors**
For many applications in analog electronics, the dependence of the FET drain current I_d on the gate voltage $V_{gs} - V_T$ is desired to be highly linear. Answer the following questions:

(a) Based on the dependence of the long-channel (drift-diffusion) and ballistic FETs, discuss which devices exhibit high linearity in the saturated on-state current regime.

(b) Show that if a sinusoidal signal of the form $V_{gs}e^{i\omega_0 t}$ is fed into the gate, then only a perfectly linear transistor characteristics will faithfully produce an amplified signal in the drain current I_d at the same frequency ω_0. Prove that any non-linearity in the I_d vs. $V_{gs} - V_T$ characteristics will produce components of drain current at harmonics of ω_0.

(c) For a perfectly linear FET, the $g_m = \partial I_d / \partial V_{gs}$ should be a constant, and all higher derivatives such as $(g_m)' = \partial^2 I_d / \partial (V_{gs})^2$, $(g_m)'' = \partial^3 I_d / \partial (V_{gs})^3$, ... must be zero. Examine whether this is the case for the long-channel drift-diffusion FET models and the ballistic FET models that are described in this chapter. If not, what factors contribute to the non-linearity?

(d) Discuss strategies by which the transistor channel design can be modified to make the FET intrinsically more linear. What bandstructures lead to higher linearity?

(e) In the chapter, we have neglected the effect of source and drain access and contact resistances on the external g_m vs. $V_{gs} - V_T$ characteristics. Argue why such resistances degrade the maximum value of the g_m, but can increase the linearity.

(f) Some design paradigms adopt a circuit-level approach with multiple V_T transistors in parallel while using the same channel semiconductor. Explore and discuss these strategies.

(19.8) **Metal-semiconductor FETs (MESFETs), junction FETs (JFETs), and so on...**
A large array of types of field-effect transistor geometries exist, with relative merits and demerits in terms of their performance, ease of fabrication. For example, the MESFET uses a metal-semiconductor Schottky gate on doped semiconductor channels. The junction-FET (JFET) uses a pn-junction gate to take advantage of large breakdown field and low leakage current of pn junctions over Schottky diodes.

Perform a survey of the various types of FETs, and summarize your results in the form of table for the following. Use typical FET metrics in the columns and different FETs in the rows for comparison. Typically a FET of a given type is attractive for certain applications –list the suitable application in one of the columns.

(19.9) **Short-term memory: MOS DRAM**

Fig. 19.30 Robert Dennard, the inventor of the 1-transistor 1-capacitor DRAM cell. He was also an original architect of the constant field scaling rule of transistors.

In a silicon MOS capacitor of Fig. 19.16, the capacitance-voltage characteristics were shown in

Fig. 19.17, and its transient behavior was discussed. The MOS capacitor is used as a dynamic random-access memory (DRAM) in modern microprocessors. Answer the following questions for exploring how the device is used for this purpose.

(a) Suppose for time $t < 0$ a p-type bulk Si MOS capacitor was in a slight accumulation regime at $V_{gs} = V_{FB} - 0.5$ Volt. Note that a p-type body implies a NMOS capacitor. If at time $t > 0$, a gate voltage $V_{gs} = V_{FB} + 1.0$ Volt is applied, and kept at that bias. Describe the physics of the time evolution of the gate capacitance.

(b) Discuss the microscopic processes that lead to, initially, a large drop in the capacitance, and ultimately in the increase to the value C_{ox}.

(c) Argue why the carriers for the inversion channel must necessarily come from the valence band by interband transition. Over what time scale can the capacitance remain at the low value?

(d) Discuss how this delay is used as the memory in DRAM. Explain why it is called "dynamic" - meaning, explain why the data desired to be stored needs to be refreshed dynamically.

(e) Very high density of DRAM memory are achieved today by using 3D capacitance geometries - discuss these innovations, and challenges in DRAM technologies in the future.

(19.10) **Moore's "law"**

Discuss the statement of Moore's law of scaling of transistors, and the fundamental quantum mechanical and practical limits encountered in scaling transistors below a few nm channel lengths. Several semiconductor materials and device innovations have been made in the past to keep the Moore's law scaling on track for several decades. Write a short essay and your view of the future of this path: and remember that several times in the past, "the news of the death of Moore's law have been greatly exaggerated"!

(19.11) **DIBL and GIDL**

Examine the following two common mechanisms that result in degradation of transistor characteristics at very short channel lengths:

(a) Drain-induced barrier lowering (DIBL): The output characteristics in the saturation regime of drain current in a FET must become independent of the drain voltage, and only be controlled by the gate voltage. In other words, the barrier height seen at the source end of the channel for electrons must only be controlled by the gate voltage, and not the drain voltage. However, for short gate lengths a large drain voltage starts competing with the gate in controlling the barrier height at the source-injection point. Discuss typical experimental indicators of DIBL and the methods used to keep it at bay.

(b) Gate-induced drain leakage (GIDL): Band-to-band tunneling currents (which are discussed in Chapter 24) are enhanced by the gate voltage in short-channel FETs on the drain side where a high field is reached. This results in gate-induced drain leakage currents. Discuss the experimental indicators and methods to counter GIDL in short-channel FETs.

(19.12) **Discrete dopant effects**

Consider a cube of silicon of side 10 μm. Calculate roughly the number of dopants in it for doping densities of $N_d = 10^{16}$ /cm^3 and $N_d = 10^{20}$ /cm^3. Now repeat the calculation for a silicon cube of side 10 nm. Discuss why since the doping density maintains the Fermi level of the semiconductor, any variation of the dopant number will lead to a variation in the important device characteristics ranging from threshold voltages to variability from one transistor to another. What methods are used to fight the discrete dopant effect in nanoscale transistors?

(19.13) **Strained silicon**

Application of strain to a semiconductor modifies its bandstructure. If done correctly, it can boost the electrical conductivity by rearranging the bands such the light effective mass carriers are populated in preference to heavier ones. Explain how strained semiconductors have been used to improve field-effect transistor performance via mobility enhancement in silicon technology.

(19.14) **HKMG: high-K, metal gate**

Discuss why in the mid 2000s the silicon CMOS industry made a transition from polysilicon gate with SiO$_2$ dielectric to high-K dielectrics and metal gates to prevent excessive gate leakage currents by tunneling. The HKMG method paved the way for the introduction of the FinFET geometry, discuss why this method is suited well for the vertical geometry and its use in the future.

(19.15) **Gain and speed of FETs**

(a) Discuss the tradeoffs between the gain and speed of field-effect transistors.

(b) Since the current gain cutoff frequency f_T is dependent on the gate length, a common metric is the $f_T \cdot L_g$ product for a semiconductor material. Argue why this product is related closely to an effective saturation velocity of carriers in the FET.

(c) The Johnson's figure-of merit (JFOM) is a product of the saturation velocity of carriers v_{sat} and the critical breakdown field F_{br}. Make a table of the JFOM for various semiconductor materials and discuss the advantages of wide-bandgap semiconductors for this metric.

(d) Make a table of the fastest FETs made of various semiconductor materials to date, including their gate lengths, contact resistance, cutoff frequencies, and breakdown voltages. Note that the fastest transistors today exceed 1 THz cutoff frequencies by virtue of high carrier velocities and low parasitic resistances and capacitances.

Fermi's Golden Rule

20

20.1 **Fermi's golden rule** 461
20.2 **Oscillating perturbations** 465
20.3 **Transitions to continuum** 467
20.4 **Kubo–Greenwood formula** 468
20.5 **Decoherence in qubits** 470
20.6 **Electron-electron scattering** 473
20.7 **Dyson series and diagrams** 475
20.8 **Zero-sum game: self energy** 476
20.9 **Chapter summary section** 477
Further reading 477
Exercises 477

In this chapter, we derive and get comfortable with Fermi's golden rule (Fig. 20.1), a central result of time-dependent perturbation theory. Because the golden rule provides a prescription to compute transition rates between quantum states, it is used heavily in subsequent chapters to understand electronic transport, and photonic properties of semiconductors. In this chapter, we learn:

- How are time-dependent quantum processes modeled?
- Under what conditions is Fermi's golden rule valid?
- How can we use the golden rule to understand electronic and photonic processes in semiconductors?
- How do time-dependent processes determine the operation of qubits in quantum computation?

Fig. 20.1 Fermi's Golden rule was initially formulated by Dirac; Fermi referred to his own version as the "Golden rule" No. 2. It is "golden" because of its wide-ranging use in evaluating rates at which quantum processes occur.

20.1 Fermi's golden rule

Consider an unperturbed quantum system in state $|\Psi_{t_0}\rangle$ at time $t = t_0$. It evolves to the state $|\Psi_t\rangle$ at a future instant t. The time evolution of the state vector is governed by the unperturbed Hamiltonian operator $\hat{H}_0$ according to the time-dependent Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}|\Psi_t\rangle = \hat{H}_0|\Psi_t\rangle. \tag{20.1}$$

If the system was in an eigenstate $|\Psi_{t_0}\rangle = |0\rangle$ of energy eigenvalue E_0 at time t_0, then the state at a future time differs from the initial state by only a phase factor:

$$\hat{H}_0|\Psi_{t_0}\rangle = E_0|\Psi_{t_0}\rangle \implies |\Psi_t\rangle = e^{-i\frac{E_0}{\hbar}(t-t_0)}|\Psi_{t_0}\rangle. \tag{20.2}$$

This is a *stationary* state; if the quantum state started in an eigentstate, it remains in that eigenstate as long as there is no perturbation. But the eigenvector still "rotates" in time with frequency $\omega_0 = E_0/\hbar$ in the Hilbert space as indicated schematically in Fig. 20.2. It is called stationary because physical observables of the eigenstate will require not the amplitude, but the inner product, which is $\langle\Psi_t|\Psi_t\rangle = \langle\Psi_{t_0}|\Psi_{t_0}\rangle$. This is manifestly stationary in time.

Now let us perturb the system with a *time-dependent* term $\hat{W}_t$. This perturbation can be due to a voltage applied on a semiconductor device, or electromagnetic waves (photons) incident on a semiconductor.

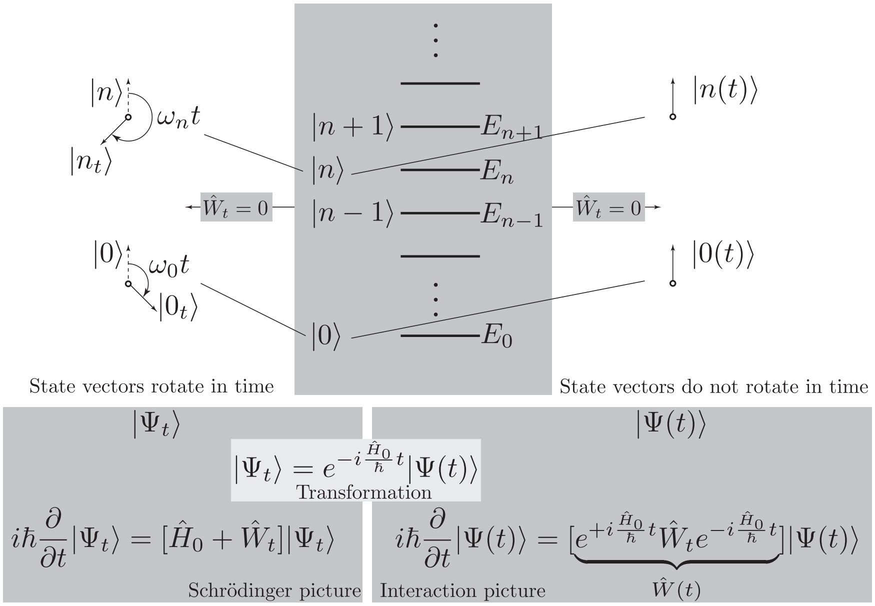

$|n\rangle$
$\omega_n t$
$|n_t\rangle$
$\hat{W}_t = 0$
$|0\rangle$
$\omega_0 t$
$|0_t\rangle$
$|n+1\rangle$
E_{n+1}
$|n\rangle$
E_n
$|n-1\rangle$
E_{n-1}
$|0\rangle$
E_0
$|n(t)\rangle$
$\hat{W}_t = 0$
$|0(t)\rangle$
State vectors rotate in time
State vectors do not rotate in time
$|\Psi_t\rangle$
$|\Psi(t)\rangle$
$|\Psi_t\rangle = e^{-i\frac{\hat{H}_0}{\hbar}t}|\Psi(t)\rangle$
Transformation
$i\hbar\frac{\partial}{\partial t}|\Psi_t\rangle = [\hat{H}_0 + \hat{W}_t]|\Psi_t\rangle$
Schrödinger picture
$i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = \underbrace{[e^{+i\frac{\hat{H}_0}{\hbar}t}\hat{W}_t e^{-i\frac{\hat{H}_0}{\hbar}t}]}_{\hat{W}(t)}|\Psi(t)\rangle$
Interaction picture

Fig. 20.2 Schrödinger vs. interaction pictures of time-evolution of quantum state.

The new Schrödinger equation for the time evolution of the state is

$$i\hbar\frac{\partial}{\partial t}|\Psi_t\rangle = [\hat{H}_0 + \hat{W}_t]|\Psi_t\rangle. \tag{20.3}$$

In principle, solving this equation will yield the complete future quantum states. In practice, this equation is unsolvable, even for the simplest of perturbations[1]. Physically, the perturbation will "scatter" a particle that was, say in state $|0\rangle$ to state $|n\rangle$. However, we had noted that even in the *absence* of perturbations, the eigenvectors were already evolving with time in the Hilbert space. For example, state vector $|0\rangle$ was rotating at an angular frequency ω_0, and state vector $|n\rangle$ at ω_n. This is shown schematically in the left of Fig. 20.2. It would be nice to work with *unperturbed* state vectors that do not change in time, as in the right of Fig. 20.2. This calls for a transformation to a vector space that "freezes" the time evolution of the unperturbed eigenvectors. Such a transformation is achieved by the relation

$$|\Psi_t\rangle = e^{-i\frac{\hat{H}_0}{\hbar}t}|\Psi(t)\rangle, \tag{20.4}$$

where $\hat{H}_0$ is the Hamiltonian *operator*. Note that the operator[2] now sits in the exponent, but it should not worry us much. We will see that it is rather useful to have it up there. The reason for this non-obvious transformation becomes clear when we put this into the form of Schrödinger equation in Equation 20.3 to get

---

[1] A two-level quantum system offers a rare and famous exactly solvable time-dependent perturbation problem, and is the basis for several practical applications ranging from NMR spectroscopy to quantum computation. It is discussed in Section 20.5.

[2] Fig. 20.2 highlights the difference between the state vectors $|\Psi_t\rangle$ and $|\Psi(t)\rangle$.

$$i\hbar\left(-\frac{i}{\hbar}\hat{H}_0 e^{-i\frac{\hat{H}_0}{\hbar}t}|\Psi(t)\rangle + e^{-i\frac{\hat{H}_0}{\hbar}t}\frac{\partial}{\partial t}|\Psi(t)\rangle\right) = [\hat{H}_0 + \hat{W}_t]e^{-i\frac{\hat{H}_0}{\hbar}t}|\Psi(t)\rangle, \tag{20.5}$$

and there is a crucial cancellation, leaving us with

$$\boxed{i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = [e^{+i\frac{\hat{H}_0}{\hbar}t}\hat{W}_t e^{-i\frac{\hat{H}_0}{\hbar}t}]|\Psi(t)\rangle = \hat{W}(t)|\Psi(t)\rangle,} \tag{20.6}$$

where $\hat{W}(t) = e^{+i\frac{\hat{H}_0}{\hbar}t}\hat{W}_t e^{-i\frac{\hat{H}_0}{\hbar}t}$. Can we take the *operator* $e^{-i\frac{\hat{H}_0}{\hbar}t}$ from the left to the right side as $e^{+i\frac{\hat{H}_0}{\hbar}t}$? Yes we can, because $e^{+i\frac{\hat{H}_0}{\hbar}t}\cdot e^{-i\frac{\hat{H}_0}{\hbar}t} = I$, the identity operator.

The boxed form of the time-evolution in Equation 20.6 is called the *interaction* picture, as opposed to the conventional form of Equation 20.3, which is called the "Schrödinger" picture. Note that if there is no perturbation, $\hat{W}_t = 0 \implies \hat{W}(t) = 0 \implies i\hbar\frac{\partial|\Psi(t)\rangle}{\partial t} = 0$. Then, $|\Psi(t)\rangle = |\Psi(t_0)\rangle$, and we have managed to find the state vector representation in which the unperturbed eigenvectors are indeed frozen in time.

Now let's turn the perturbation $\hat{W}_t$ on. Formally, the state vector at time t in the interaction representation is obtained by integrating both sides of Equation 20.6:

$$|\Psi(t)\rangle = |\Psi(t_0)\rangle + \frac{1}{i\hbar}\int_{t_0}^{t} dt'\hat{W}(t')|\Psi(t')\rangle, \tag{20.7}$$

and it looks as if we have solved the problem. However, there is a catch – the unknown state vector $|\Psi(t)\rangle$ appears also on the right side – inside the integral. This is a recursive relation! It reminds of the Brillouin-Wigner form of non-degenerate perturbation theory. Let's try to iterate the formal solution once:

$$|\Psi(t)\rangle = |\Psi(t_0)\rangle + \frac{1}{i\hbar}\int_{t_0}^{t} dt'\hat{W}(t')\left[|\Psi(t_0)\rangle + \frac{1}{i\hbar}\int_{t_0}^{t'} dt''\hat{W}(t'')|\Psi(t'')\rangle\right], \tag{20.8}$$

and then keep going:

$$|\Psi(t)\rangle = \underbrace{|\Psi(t_0)\rangle}_{\sim W^0} + \underbrace{\frac{1}{i\hbar}\int_{t_0}^{t} dt'\hat{W}(t')|\Psi(t_0)\rangle}_{\sim W^1} + \underbrace{\frac{1}{(i\hbar)^2}\int_{t_0}^{t} dt'\hat{W}(t')\int_{t_0}^{t'} dt''\hat{W}(t'')|\Psi(t_0)\rangle}_{\sim W^2} + ... \tag{20.9}$$

We thus obtain a formal perturbation series to many orders. The hope is that the series converges rapidly if the perturbation is "small", because successive terms increase as a power law, which for a small number gets even smaller. Let's accept that weak argument now at face value, and we return later to address, justify, and where possible, fix this cavalier approximation.

Let $|\Psi(t_0)\rangle = |0\rangle$ be the initial state of the quantum system. The perturbation is turned on at time t_0. The probability *amplitude* for the system to be found in state $|n\rangle$ at time $t(> t_0)$ is $\langle n|\Psi_t\rangle$. Note the Schrödinger representation! But the transformation from Schrödinger to interaction picture helps: $\langle n|\Psi_t\rangle = \langle n|e^{-i\frac{\hat{H}_0}{\hbar}t}\Psi(t)\rangle = e^{-i\frac{E_n}{\hbar}t}\langle n|\Psi(t)\rangle$. This implies $|\langle n|\Psi_t\rangle|^2 = |\langle n|\Psi(t)\rangle|^2$ – for all eigenstates $|n\rangle$. Let us make an approximation in this section and retain only the first order term in the perturbation series. We will return later and discuss the higher order terms that capture multiple-scattering events. Retaining only the terms of Eq. 20.9 to first order in the perturbation $\hat{W}$ gives

$$\langle n|\Psi(t)\rangle \approx \underbrace{\langle n|0\rangle}_{=0} + \frac{1}{i\hbar}\int_{t_0}^{t} dt' \langle n|W(t')|0\rangle = \frac{1}{i\hbar}\int_{t_0}^{t} dt' \langle n|e^{+i\frac{\hat{H}_0}{\hbar}t'} W_{t'} e^{-i\frac{\hat{H}_0}{\hbar}t'}|0\rangle. \tag{20.10}$$

Let us assume the perturbation to be of the form $\hat{W}_t = e^{\eta t}W$ representing a "slow turn on", with $\eta = 0^+$, and $W = W(r)$ a function that depends only on space (Fig. 20.3). If $\eta = 0$, then the perturbation is time-independent. But if $\eta = 0^+$, then $e^{\eta t_0} \to 0$ as $t_0 \to -\infty$. This construction thus effectively kills the perturbation far in the distant past, but slowly turns it on to full strength at $t = 0$. We will discuss more of the physics buried inside η later. For now, we accept it as a mathematical construction, with the understanding that we will take the limit $\eta \to 0$ at the end. Then, the amplitude in state $|n\rangle$ simplifies:

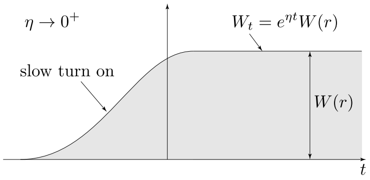


Fig. 20.3 The perturbation potential is turned on slowly to full strength.

$$\begin{aligned}\langle n|\Psi(t)\rangle &\approx \frac{1}{i\hbar}\int_{t_0}^{t} dt' \underbrace{\boxed{\langle n|e^{+i\frac{\hat{H}_0}{\hbar}t'}}}_{e^{+i\frac{E_n}{\hbar}t'}\langle n|} e^{\eta t'} W \underbrace{\boxed{e^{-i\frac{\hat{H}_0}{\hbar}t'}|0\rangle}}_{e^{-i\frac{E_0}{\hbar}t'}|0\rangle} \\ &= \frac{\langle n|W|0\rangle}{i\hbar}\int_{t_0}^{t} dt' e^{i\left(\frac{E_n-E_0}{\hbar}\right)t'} e^{\eta t'},\end{aligned} \tag{20.11}$$

and the integral over time may be evaluated exactly to yield

$$\int_{t_0}^{t} dt' e^{i\left(\frac{E_n-E_0}{\hbar}\right)t'} e^{\eta t'} = \frac{e^{i\left(\frac{E_n-E_0}{\hbar}\right)t} e^{\eta t} - e^{i\left(\frac{E_n-E_0}{\hbar}\right)t_0} e^{\eta t_0}}{i\left(\frac{E_n-E_0}{\hbar}\right)+\eta} \underbrace{=}_{t_0 \to -\infty} \frac{e^{i\left(\frac{E_n-E_0}{\hbar}\right)t} e^{\eta t}}{i\left(\frac{E_n-E_0}{\hbar}\right)+\eta}. \tag{20.12}$$

The amplitude then is

$$\langle n|\Psi(t)\rangle \approx \frac{\langle n|W|0\rangle}{i\hbar} \cdot \frac{e^{i\left(\frac{E_n-E_0}{\hbar}\right)t} e^{\eta t}}{i\left(\frac{E_n-E_0}{\hbar}\right)+\eta} = \langle n|W|0\rangle \cdot \frac{e^{i\left(\frac{E_n-E_0}{\hbar}\right)t} e^{\eta t}}{(E_0-E_n)+i\hbar\eta}. \tag{20.13}$$

The *probability* of the state making a transition from $|0\rangle$ to $|n\rangle$ at time

t is

$$|\langle n|\Psi_t\rangle|^2 = |\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \frac{e^{2\eta t}}{(E_0 - E_n)^2 + (\hbar\eta)^2}. \tag{20.14}$$

The *rate* of transitions from state $|0\rangle \to |n\rangle$ is

$$\frac{1}{\tau_{|0\rangle\to|n\rangle}} = \frac{d}{dt}|\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \left(\frac{2\eta}{(E_0 - E_n)^2 + (\hbar\eta)^2}\right) e^{2\eta t}. \tag{20.15}$$

Now we take $\eta \to 0^+$. The third term $e^{2\eta t} \to 1$, but we must be careful with the quantity in the round brackets. When $\eta \to 0$, this quantity is 0, *except* when the term $E_0 - E_n = 0$. For $E_0 = E_n$, the term appears to be in the $0/0$ indeterminate form. By making a plot of this function, we can convince ourselves that it approaches a Dirac delta function $\delta(...)$ in the variable $E_0 - E_n$. The mathematical identity $\lim_{\eta\to 0^+} \frac{2\eta}{x^2+\eta^2} = \lim_{\eta\to 0^+} \frac{1}{i}[\frac{1}{x-i\eta} - \frac{1}{x+i\eta}] = 2\pi\delta(x)$ confirms this: in the limit, the term indeed becomes the Dirac delta function.

Then, using $\delta(ax) = \delta(x)/|a|$, the rate of transitions is given by

$$\boxed{\frac{1}{\tau_{|0\rangle\to|n\rangle}} \approx \frac{2\pi}{\hbar}|\langle n|W|0\rangle|^2 \delta(E_0 - E_n),} \tag{20.16}$$

which is the Fermi's golden rule. The general form of the transition rate is $2\pi/\hbar$ times the transition matrix element squared, times a Dirac delta function as a statement of energy conservation[3].

20.2 Oscillating perturbations

Now suppose the perturbation potential was oscillating in time. Such cases are encountered for an electron interacting with photons, or phonons, both of which by virtue of being waves, present time-varying scattering potentials for the electron. The mathematical nature of such oscillating perturbations with a slow turn-on is

$$W_t = 2We^{\eta t}\cos(\omega t) = e^{\eta t}W(e^{i\omega t} + e^{-i\omega t}), \tag{20.17}$$

which leads to a $|0\rangle \to |n\rangle$ transition amplitude

$$\langle n|\Psi(t)\rangle \approx \frac{\langle n|W|0\rangle}{i\hbar}\left(\int_{t_0}^{t} dt' e^{i\left(\frac{E_n - E_0 + \hbar\omega}{\hbar}\right)t'} e^{\eta t'} + \int_{t_0}^{t} dt' e^{i\left(\frac{E_n - E_0 - \hbar\omega}{\hbar}\right)t'} e^{\eta t'}\right). \tag{20.18}$$

Similar to Equations 20.12 and 20.13, evaluating the integral with $t_0 \to -\infty$, we get the amplitude for transitions

$$\langle n|\Psi(t)\rangle \approx \langle n|W|0\rangle \cdot \left(\frac{e^{i\left(\frac{E_n - E_0 + \hbar\omega}{\hbar}\right)t} e^{\eta t}}{(E_0 - E_n + \hbar\omega) + i\hbar\eta} + \frac{e^{i\left(\frac{E_n - E_0 - \hbar\omega}{\hbar}\right)t} e^{\eta t}}{(E_0 - E_n - \hbar\omega) + i\hbar\eta}\right). \tag{20.19}$$

---

[3] See exercises of this chapter for a collection of solved examples, and unsolved ones for you to practice. Applications of time-dependent perturbation theory abound in semiconductor physics and are found generously sprinkled throughout following chapters.

The probability of transition is then

$$|\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \cdot \Big[\frac{e^{2\eta t}}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} + \frac{e^{2\eta t}}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2} + \frac{e^{2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega+i\hbar\eta)(E_0-E_n-\hbar\omega-i\hbar\eta)} + \frac{e^{-2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega-i\hbar\eta)(E_0-E_n-\hbar\omega+i\hbar\eta)}\Big], \tag{20.20}$$

and the rate of transition is then

$$\frac{d}{dt}|\langle n|\Psi(t)\rangle|^2 \approx |\langle n|W|0\rangle|^2 \times \Big[\frac{2\eta e^{2\eta t}}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} + \frac{2\eta e^{2\eta t}}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2} + \frac{2(\eta+i\omega)e^{2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega+i\hbar\eta)(E_0-E_n-\hbar\omega-i\hbar\eta)} + \frac{2(\eta-i\omega)e^{-2i\omega t}e^{2\eta t}}{(E_0-E_n+\hbar\omega-i\hbar\eta)(E_0-E_n-\hbar\omega+i\hbar\eta)}\Big]. \tag{20.21}$$

Notice that the last two (interference) terms are a complex conjugate pair, which they must be, because the rate of transition is real. The sum is then 2× the real part of either term. After some manipulations, we get

$$\frac{d}{dt}|\langle n|\Psi(t)\rangle|^2 \approx \langle n|W|0\rangle|^2 e^{2\eta t} \cdot \left(\frac{2\eta}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} + \frac{2\eta}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2}\right)[1-\cos(2\omega t)] + 2\sin(2\omega t)\left(\frac{E_0-E_n+\hbar\omega}{(E_0-E_n+\hbar\omega)^2+(\hbar\eta)^2} - \frac{E_0-E_n-\hbar\omega}{(E_0-E_n-\hbar\omega)^2+(\hbar\eta)^2}\right). \tag{20.22}$$

Note that the rate has a part that does not oscillate, and another which does, with *twice* the frequency of the perturbing potential. If we average over a few periods of the oscillation, $\langle\cos(2\omega t)\rangle_t = \langle\sin(2\omega t)\rangle_t = 0$. Then, by taking the limit $\eta \to 0^+$ in the same fashion as in Equation 20.16, we obtain the Fermi's golden rule for oscillating perturbations:

$$\boxed{\frac{1}{\tau_{|0\rangle\to|n\rangle}} \approx \frac{2\pi}{\hbar} \times |\langle n|W|0\rangle|^2 \times [\underbrace{\delta(E_0-E_n+\hbar\omega)}_{\text{absorption}} + \underbrace{\delta(E_0-E_n-\hbar\omega)}_{\text{emission}}].} \tag{20.23}$$

The Dirac delta functions now indicate that the exchange of energy between the quantum system and the perturbing field is through *quanta* of energy: either by absorption, leading to $E_n = E_0 + \hbar\omega$, or emission, leading to $E_n = E_0 - \hbar\omega$. The rates of absorption and emission are the same. Which process (emission or absorption) dominates depends on the occupation functions of the quantum states[4].

[4] Typical transition rates for electrons due to interaction with light (photons) is $1/\tau \sim 10^9$/s, or it takes $\sim$ ns for an optical transition. The same for phonons is $1/\tau \sim 10^{12}$/s; it takes $\sim$ps for thermal transitions. These processes are discussed in following chapters.

20.3 Transitions to continuum

The Fermi golden rule results of Equations 20.16 and 20.23 are in forms that are suitable for tracking transitions between *discrete*, or individual states $|0\rangle$ and $|n\rangle$. For many situations encountered in semiconductors, these transitions will be between states within, or between energy bands, where a *continuum* of states exist. In those cases, the *net* transition rate will be obtained by summing over all relevant states. Even the transition between manifestly discrete states – for example the electron ground state of hydrogen atom to the first excited state – by the absorption of a photon – occurs by the interaction between the discrete electron states and the states of the electromagnetic spectrum, which form a continuum.

As an example, consider the transitions between electron states in the conduction band due to a point scatterer in a 3D semiconductor (see Fig. 20.4). Let us say the point scatterer potential is $W(r) = V_0\delta(\mathbf{r})$, with V_0 in units of eV·m^3. This is not an oscillating potential, so we use the golden rule result of Equation 20.16. We first find the matrix element between plane-wave electron states $|\mathbf{k}\rangle$ and $|\mathbf{k}'\rangle$:

$$\langle\mathbf{k}'|V_0\delta(\mathbf{r})|\mathbf{k}\rangle = \int d^3\mathbf{r}\left(\frac{e^{-i\mathbf{k}'\cdot\mathbf{r}}}{\sqrt{V}}\right)V_0\delta(\mathbf{r})\left(\frac{e^{+i\mathbf{k}\cdot\mathbf{r}}}{\sqrt{V}}\right) = \frac{V_0}{V}. \qquad (20.24)$$

Here we have used the property that the Fourier transform of a Dirac delta function is equal to 1. Notice that the matrix element for scattering of plane waves is the Fourier-transform of the scattering potential, a property referred to as the **Born approximation**. The physical meaning is that the perturbation potential $W(r)$ can scatter electrons from a state $\mathbf{k}$ to a state $\mathbf{k}^{'}$ only if its Fourier transform $W_{\mathbf{q}} = \frac{1}{V}\int d^3\mathbf{r}e^{i\mathbf{q}\cdot\mathbf{r}}W(r)$ has a non-zero coefficient at $\mathbf{q} = \mathbf{k} - \mathbf{k}^{'}$. Then, the transition (or scattering) rate to any final state $|\mathbf{k}'\rangle$ is

$$\frac{1}{\tau(|\mathbf{k}\rangle \to |\mathbf{k}'\rangle)} = \frac{2\pi}{\hbar}\left(\frac{V_0}{V}\right)^2\delta(E_{\mathbf{k}} - E_{\mathbf{k}'}). \qquad (20.25)$$

The net scattering "out" of state $|\mathbf{k}\rangle$ into the continuum of states $|\mathbf{k}'\rangle$ is then given by

$$\frac{1}{\tau(|\mathbf{k}\rangle)} = \sum_{\mathbf{k}'}\frac{1}{\tau(|\mathbf{k}\rangle \to |\mathbf{k}'\rangle)} = \frac{2\pi}{\hbar}\left(\frac{V_0}{V}\right)^2\underbrace{\sum_{\mathbf{k}'}\delta(E_{\mathbf{k}} - E_{\mathbf{k}'})}_{G(E_{\mathbf{k}})}, \qquad (20.26)$$

where we note that the sum over final states of the Dirac delta function is the density of states[5] $G(E_{\mathbf{k}})$ in units eV^{-1} of the electron at energy $E_{\mathbf{k}}$. This procedure illustrates an important result – the scattering rate from a single initial state into a continuum of states is **proportional to the density of available final states**. The strength of scattering increases as the square of the scattering potential. The occurrence of

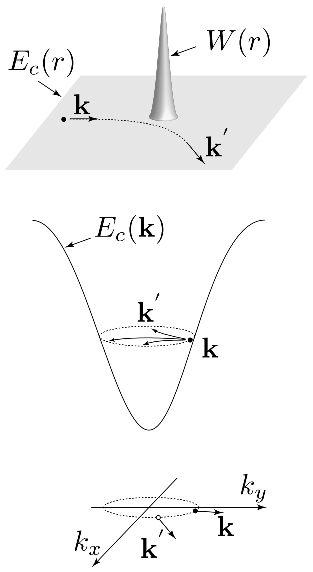


Fig. 20.4 Illustration of scattering of a conduction electron by a perturbation potential $W(r)$ in real space, and the same process illustrated in the $\mathbf{k}$-space for the electron bandstructure. For an elastic scattering event, there are a continuum of available final states.

---

[5] Note here that $G(E_{\mathbf{k}})$ must *not* include the spin-degeneracy for scattering events that cannot flip the spin of an electron.

the (volume)^2 term in the denominator may be disconcerting at first. However, the macroscopic volume (or area, or length) terms will for most cases cancel out because of purely physical reasons. For example, for the problem illustrated here, if instead of just *one* point scatterer, we had N, the *density* of scatterers is $n_{sc} = N/V$. Together with the conversion $\sum_{\mathbf{k}'} \to V \int d^3\mathbf{k}'/(2\pi)^3$, we obtain

$$\frac{1}{\tau(E_{\mathbf{k}})} = \frac{2\pi}{\hbar}\left(\frac{V_0}{V}\right)^2 n_{sc} V \int \frac{d^3\mathbf{k}'}{\frac{(2\pi)^3}{V}} \delta(E_{\mathbf{k}} - E_{\mathbf{k}'}) = \frac{2\pi}{\hbar} V_0^2 n_{sc} g(E_{\mathbf{k}}). \tag{20.27}$$

Here the density of states $g(E_{\mathbf{k}})$ is per unit volume, in units 1/(eV.m^3), as is standard in semiconductor physics. The scattering rate is linearly proportional to the density of scatterers. What is not immediately clear is how can we capture the effect of N scatterers by just multiplying the individual scatterer rate by N. This can be done if the scatterers are uncorrelated, as will be discussed in Chapter 23. For now, note that the macroscopic volume has canceled out, as promised[6].

20.4 Kubo–Greenwood formula

Fermi's golden rule helps find the electronic response of any solid to an oscillating electric field $F\cos(\omega t)$ as shown in Fig. 20.5. Electrons in the solid respond to the electric field by carrying a current that is linearly proportional to the field, with $\sigma(\omega)$ the electronic conductivity as the proportionality constant. The rate of energy absorbed and dissipated in the solid is $\frac{1}{2}\sigma(\omega)F^2$. This must be due to the difference of downward and upward electronic transition processes as indicated in the figure. Using this idea, the Kubo–Greenwood formula expresses the macroscopic electronic conductivity $\sigma(\omega)$ of the solid in terms of the microscopic electronic density of states $g(E)$.

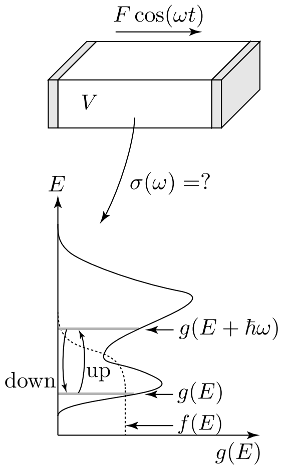


Fig. 20.5 The linear response of a solid subjected to an oscillating electric field $F\cos(\omega t)$ is its electronic conductivity $\sigma(\omega)$. The Kubo–Greenwood formula relates $\sigma(\omega)$ to the electronic density of states $g(E)$ of the solid by the application of Fermi's golden rule.

The method hinges on energy conservation in the transitions. Consider the density of states $g(E)$ of the allowed electron eigenvalues in the solid, with an occupation function $f(E)$. Writing the electric field as $\frac{1}{2}F(e^{i\omega t} + e^{-i\omega t})$, the perturbation potential experienced by the electrons is $W(x,t) = \frac{1}{2}qFx(e^{i\omega t} + e^{-i\omega t})$. Using the golden rule (Equation 20.23), the rate of an upward transition from state $|i\rangle \to |f\rangle$ is

$$\begin{aligned}\frac{1}{\tau_{up}(i \to f)} &= \frac{2\pi}{\hbar}|\langle f|\frac{1}{2}qFx|i\rangle|^2 \delta(E_f - (E_i + \hbar\omega)) \\ &= \frac{\pi q^2 F^2}{2\hbar}|\langle f|x|i\rangle|^2 \delta(E_f - (E_i + \hbar\omega)).\end{aligned} \tag{20.28}$$

The matrix element $\langle f|x|i\rangle$ is dimensionally a length, and is called a **dipole matrix element**. It is sometimes preferable to trade off x with the derivative $\frac{\partial}{\partial x}$ to get the closely related **momentum matrix element** $\langle f| - i\hbar\frac{\partial}{\partial x}|i\rangle$. This is achieved using the property $[x, \hat{H}_0] = \frac{\hbar^2}{m_e}\frac{\partial}{\partial x}$ which was derived in Equation 9.38 for an unperturbed Hamiltonian $\hat{H}_0$. Because the initial states are eigenstates of the unperturbed Hamiltonian

---

[6] For typical scattering center densities of $n_{sc} \sim 10^{17}/\text{cm}^3$, the scattering rate is of the order of $1/\tau \sim 10^{12}/\text{s}$, or a scattering event occurs every ps. Of course this depends on the scattering potential and other parameters, and there are also several types of scattering rates, as is investigated in following chapters on Boltzmann transport (Chapter 21), phonons (Chapter 22), and mobility (Chapter 23).

$\hat{H}_0|i\rangle = E_i|i\rangle$ and $\hat{H}_0|f\rangle = E_f|f\rangle$, this leads to

$$\langle f|\frac{\partial}{\partial x}|i\rangle = \frac{m_e}{\hbar^2}\langle f|[x,\hat{H}]|i\rangle = \frac{m_e}{\hbar^2}\langle f|x\hat{H}-\hat{H}x|i\rangle$$

$$= \frac{m_e}{\hbar^2}(\langle f|x\underbrace{\hat{H}|i\rangle}_{E_i|i\rangle} - \underbrace{\langle f|\hat{H}}_{E_f\langle f|}x|i\rangle) = \frac{m_e(E_i-E_f)}{\hbar^2}\langle f|x|i\rangle$$

$$E_f = E_i+\hbar\omega \implies \langle f|x|i\rangle = -\frac{\hbar}{m_e\omega}\langle f|\frac{\partial}{\partial x}|i\rangle. \tag{20.29}$$

Writing the matrix element as $D = \langle f|\frac{\partial}{\partial x}|i\rangle = \int d^3x \cdot \psi_{E+\hbar\omega}\frac{\partial}{\partial x}\psi_E$,

$$\frac{1}{\tau_{up}(i\to f)} = \frac{\pi q^2\hbar}{2m_e^2\omega^2}F^2|D|^2\delta(E_f-(E_i+\hbar\omega))$$

$$\frac{1}{\tau_{up}} = \sum_f \frac{1}{\tau_{up}(i\to f)} = \frac{\pi q^2\hbar}{2m_e^2\omega^2}F^2\sum_f|D|^2\delta(E_f-(E_i+\hbar\omega))$$

$$\frac{1}{\tau_{up}} = \frac{\pi q^2\hbar V}{2m_e^2\omega^2}\cdot F^2|D|^2\cdot g(E+\hbar\omega) \text{ and } \frac{1}{\tau_{down}} = \frac{\pi q^2\hbar V}{2m_e^2\omega^2}\cdot F^2|D|^2\cdot g(E). \tag{20.30}$$

Note that the net upward transitions from a single state $E_i = E$ to all possible final states $E_f = E+\hbar\omega$ is proportional to the number of final states $g(E+\hbar\omega)V$ in a unit energy interval. For the same reason, the rate of downward transitions is proportional to $g(E)V$. For energy conservation, we must also sum over all possible initial states. Because of the Pauli exclusion principle, the transitions actually occur if the initial states are filled and final states are empty. Energy conservation per unit time then gives us the relation

$$\frac{1}{2}\sigma(\omega)F^2 = P_{up}-P_{down}, \text{ where}$$

$$P_{up} = 2\cdot\hbar\omega\cdot\int dE\cdot\frac{1}{\tau_{up}}\cdot f(E)[1-f(E+\hbar\omega)]\cdot g(E), \text{ and}$$

$$P_{down} = 2\cdot\hbar\omega\cdot\int dE\cdot\frac{1}{\tau_{down}}\cdot f(E+\hbar\omega)[1-f(E)]\cdot g(E+\hbar\omega)$$

$$P_{up}-P_{down} = 2\cdot\hbar\omega\cdot\frac{\pi q^2\hbar V}{2m_e^2\omega^2}\cdot F^2\cdot\int dE|D|^2\cdot$$

$$\Big(f(E)(1-f(E+\hbar\omega))-f(E+\hbar\omega)(1-f(E))\Big)\cdot g(E)g(E+\hbar\omega)$$

$$\frac{1}{2}\sigma(\omega)F^2 = \frac{\pi q^2\hbar^3 V}{m_e^2}\cdot F^2\cdot\int dE|D|^2 g(E)g(E+\hbar\omega)[\frac{f(E)-f(E+\hbar\omega)}{\hbar\omega}]$$

$$\implies \boxed{\sigma(\omega) = \frac{2\pi q^2\hbar^3 V}{m_e^2}\int dE|D|^2 g(E)g(E+\hbar\omega)[\frac{f(E)-f(E+\hbar\omega)}{\hbar\omega}].} \tag{20.31}$$

The final boxed form is the **Kubo–Greenwood formula**. It provides a method to compute the electronic conductivity $\sigma(\omega)$ from the electronic density of states $g(E)$ in the solid, and the occupation function $f(E)$. As an example, for near DC excitation $\omega \to 0$, the term in the square brackets becomes $-\frac{\partial f}{\partial E}$, the derivative of the occupation function. If the occupation function follows the Fermi–Dirac distribution, at low temperatures $-\frac{\partial f}{\partial E} \approx \delta(E - E_F)$, the electronic conductivity becomes $\sigma(0) \approx \frac{2\pi q^2 \hbar^3 V}{m_e^2} |D|^2_{av} g(E_F)^2$. N. Mott and D. J. Thouless have shown that for normal band-transport of extended electronic states the Kubo–Greenwood formula yields the well-known electrical conductivity $\sigma = \frac{nq^2\tau}{m^\star}$ as discussed in Chapter 21 on Boltzmann transport. Exercise 20.9 takes you through this derivation[7].

But the utility of this formula extends to cases where Boltzmann transport fails, such as in disordered or amorphous solids, in the presence of localization, and also as a unified response to both DC and high frequency fields. For example, $\sigma(\omega)$ can also model interband optical absorption. Surprisingly, even conductivity resulting from very high electric fields such as interband tunneling in semiconductors can be computed using this formula in far more generality than several other approaches[8]. The Kubo–Greenwood formula is an example of **fluctuation-dissipation theorems** of statistical mechanics that relate the microscopic fluctuations or "noise" quantitatively to energy dissipation. Another example is the Einstein relation $\frac{D}{\mu} = \frac{k_b T}{q}$, which relates the fluctuations in diffusive motion via the diffusion constant D to the dissipation due to drift via the mobility μ.

20.5 Decoherence in qubits

In Chapter 5 we discussed a few exact solutions of the Schrödinger equation in time-independent potentials, such as the particle in a box, the harmonic oscillator, and the hydrogen atom. Though exact solutions are rare in time-dependent perturbation theory, there is one problem of tremendous practical importance where such a solution does exist: **a 2-level quantum system in an oscillating electric field** (see Fig. 20.6). This problem has been a source of a stream of new ideas in the past century. The latest practical manifestation of this system is as a quantum bit, or a **qubit**.

Fig. 20.6 Isidore Rabi pioneered magnetic resonance and molecular beam techniques. He developed the exact solutions of the 2-level quantum systems and the oscillations are named after him. Rabi was awarded the 1944 Nobel Prize in Physics.

The time evolution of a quantum state formed by the superposition $|\Psi(t)\rangle = \sum_n c_n(t)|n\rangle$ of a large number of states in the interaction picture is given by Equation 20.6 as

$$i\hbar \frac{d}{dt}\begin{bmatrix} c_1(t) \\ c_2(t) \\ c_3(t) \\ \vdots \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12}e^{i\omega_{12}t} & W_{13}e^{i\omega_{13}t} & \cdots \\ W_{21}e^{i\omega_{21}t} & W_{22} & W_{23}e^{i\omega_{23}t} & \cdots \\ W_{31}e^{i\omega_{31}t} & W_{32}e^{i\omega_{32}t} & W_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} c_1(t) \\ c_2(t) \\ c_3(t) \\ \vdots \end{bmatrix}. \tag{20.32}$$

The matrix defines coupled differential equations that correlate all coefficients while preserving $\sum_n |c_n(t)|^2 = 1$ at all times. For a time-

---

[7] It is important to note that the Kubo–Greenwood formula provides a path to understanding the electrical conductivity of any solid for which the DOS $g(E)$ is known. An example is amorphous solids, for which the whole framework of periodicity and Bloch theorem fails. For example this formalism explains why amorphous SiO_2 is insulating.

[8] Chapter 24, Section 24.15 discusses the Kubo formalism for tunneling transport.

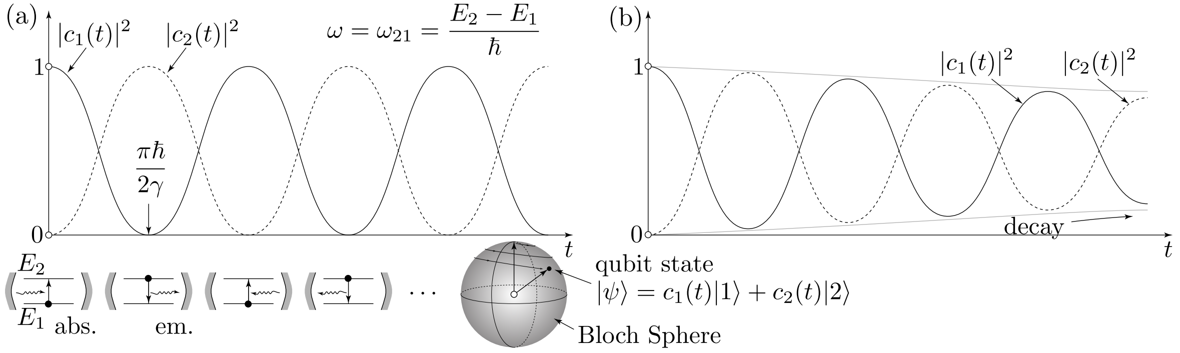


Fig. 20.7 (a) Rabi oscillations in a 2-level quantum system. It models a qubit, which could be formed from 2 discrete electron eigenvalues in a semiconductor quantum dot, and a photon trapped between mirrors. (b) If either the 2 levels are not decoupled from the crystal, or if the photon escapes the cavity, the qubit state decays and loses its quantum information processing capacity.

dependent perturbation $\hat{W}(t)$, the W_{nm} in the matrix are also time-dependent. For a 2-level quantum system, the matrix reduces to

$$i\hbar\frac{d}{dt}\begin{bmatrix} c_1(t) \\ c_2(t) \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12}e^{i\omega_{12}t} \\ W_{21}e^{i\omega_{21}t} & W_{22} \end{bmatrix}\begin{bmatrix} c_1(t) \\ c_2(t) \end{bmatrix}. \tag{20.33}$$

For the two unperturbed levels, $\hat{H}_0|1\rangle = E_1|1\rangle$ and $\hat{H}_0|2\rangle = E_2|2\rangle$. The perturbing potential from the oscillating electric field is $W(r,t) = qE_0x(e^{i\omega t} + e^{-i\omega t})$. $|W_{12}| = |\langle 2|qE_0x|1\rangle| = \gamma$ is the magnitude of the matrix element. Let the 2-level system have one electron. If at time $t = 0$ state $|1\rangle$ is occupied, and state $|2\rangle$ is empty, $c_1(0) = 1$ and $c_2(0) = 0$. In Exercise 20.10, you will show that then, the time evolution of the states is characterized by **Rabi oscillations** with probabilities

$$\boxed{|c_2(t)|^2 = C^2\sin^2\Omega t} \text{ and } \boxed{|c_1(t)|^2 = 1 - C^2\sin^2\Omega t} = 1 - |c_2(t)|^2,$$

$$\text{where } \boxed{C^2 = \frac{(\frac{\gamma}{\hbar})^2}{(\frac{\gamma}{\hbar})^2 + (\frac{\omega-\omega_{21}}{2})^2}} \text{ and } \boxed{\Omega = \sqrt{(\frac{\gamma}{\hbar})^2 + (\frac{\omega-\omega_{21}}{2})^2}}. \tag{20.34}$$

Fig. 20.7 shows the Rabi oscillations of the qubit. The electron absorbs the energy of the oscillating electric field (the photon) and transitions to the upper state in the first half of a cycle. If the photon energy $\hbar\omega$ is "tuned" to the energy level difference $\hbar\omega_{21} = \frac{E_2 - E_1}{\hbar}$, the transition can be complete because $C^2 = 1$, and $|c_2(t)|^2$ can reach unity as seen in Figure 20.7 (a). In the next half of the cycle, the electron makes a downward transition and it emits the photon back. This cycle of absorption and emission is repeated. The superposition state vector for the qubit $|\psi\rangle = c_1(t)|1\rangle + c_2(t)|2\rangle$ is pictured as a vector on a Bloch sphere. In Rabi oscillations, the tip of this state vector traverses a spiral trajectory going from the north to south pole, and back in a continuous fashion.

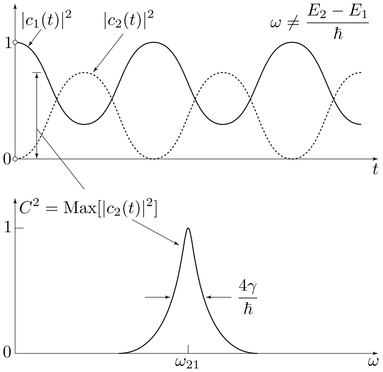


Fig. 20.8 Rabi resonance quality.

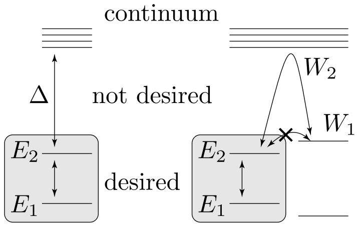


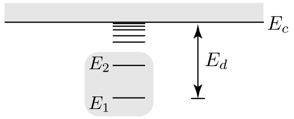
Fig. 20.9 Decoherence of qubits due to the effect of environment and energetically close eigenvalues. That even topological protection does not make qubits immune to decoherence is explained by Fermi's golden rule.

If the photon energy is ”detuned” to $\omega \neq \omega_{21} = \frac{E_2 - E_1}{\hbar}$, $C^2 < 1$, and the transition cannot be complete, as shown in Fig. 20.8. The oscillations still occur, but the state vector only spans a part of the Bloch sphere before returning to where it started. The extent of this detuning and the quality of the Rabi resonance is given by the full-width at half maximum of $4\gamma/\hbar$. This depends on the strength of the electric field of the oscillation: a stronger field leads to a sharper resonance, and near-complete transitions become possible even with slight detuning.

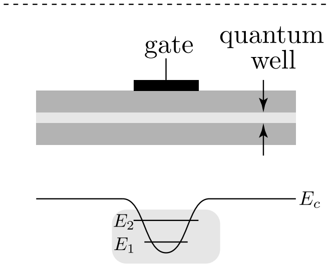
Why is this 2-level system called a **qubit**, and how is it different from a classical bit? A classical bit is composed of only two states: $|1\rangle$ and $|2\rangle$, which we may call a zero and a one. The tip of the state vector for a classical bit accesses only two points on the surface of the Bloch sphere: the north pole, and the south pole. But in the qubit, superposition of the states allows the tip of the state vector to traverse the entire surface of the Bloch sphere, opening up a much larger space to encode information than the classical bit. Combined with the ability to entangle multiple qubits by shuttling photons between them opens up the possibility to do certain computations that are difficult with classical bits. This is the promise of **quantum computation**. Though conceptually simple, the technical challenges are: 1) preventing dephasing and decoherence in the time evolution, and 2) treading a fine line between letting the system evolve in time untouched, and also accessing its state at intervals to extract useful computation from it.

The coherent oscillations may be compromised if the energies E_2 and E_1 fluctuate, say due to vibrations or coupling of the state hosting the two electronic states to an environment as shown in Fig. 20.9. The photons can leak out of the ”mirrors” meant to confine them. These processes can lead to the loss of coherence of the state while still maintaining $|c_1(t)|^2 + |c_2(t)|^2 = 1$, which is pictured as the tip of the state vector undergoing a jittery, diffusive random-walk type of motion on the Bloch sphere instead of the well-defined coherent spirals. Or they may actually cause a decay of the amplitudes themselves, leading to $|c_1(t)|^2 + |c_2(t)|^2 = e^{-\Gamma t}$. This may happen for example if the electron itself escapes into the surroundings, at a rate $1/\Gamma$, shrinking the amplitude to zero. Such decays and sources of decoherence are shown schematically in Fig. 20.9. Fermi's golden rule is used to compute and reduce the undesired transitions and dephasing via coupling of the desired states for qubits, to the undesired continuum of states in the environment in qubit design. Even if some energy eigenvalues close to the qubits may be ”topologically protected”, meaning their direct matrix elements are zero preventing direct transitions, indirect higher-order transitions mediated by the continuum are still possible.

Semiconductor structures are at the forefront in the practical realization of qubits for quantum computation. Fig. 20.10 indicates early realizations. Recall from Chapter 14 that the extra electron tied to a donor atom in a semiconductor behaves as a hydrogen atom by the effective mass theory. The states of the donor can be used for a



Donor atom qubit



Quantum dot qubit

Fig. 20.10 Practical realizations of qubits using semiconductors. So that thermal vibrations do not cause unintended transitions between the states E_1 and E_2, qubits are operated at temperatures $k_b T << E_2 - E_1$.

donor qubit, with a microwave photon driving the Rabi oscillations between the two lowest states. The other states must be made inaccessible, which is ensured by cooling to cryogenic temperatures such that $k_b T << \Delta$. A **quantum-dot qubit** may be fabricated using an electrostatic potential well created in the conduction band of a quantum well. While the quantum well confines electrons in the vertical direction, the nanoscale gate confines it in the plane, allowing the formation of a gate-controlled quantum dot. The gate control also allows loading and unloading electrons into the dot. The transitions may be driven by microwave photons. The different qubit states may be the different spin states of the electron, in which case it is called a **spin qubit**. Several other platforms are under investigation for qubits, ranging from superconducting junctions to ion traps. Though not without its challenges, semiconductor-based qubits stand to benefit from the enormous progress brought about by the development of the physics, materials, and precision nanoscale fabrication techniques that produce the ~ 5 nm sized billion-transistor chips of modern silicon electronics.

20.6 Electron-electron scattering

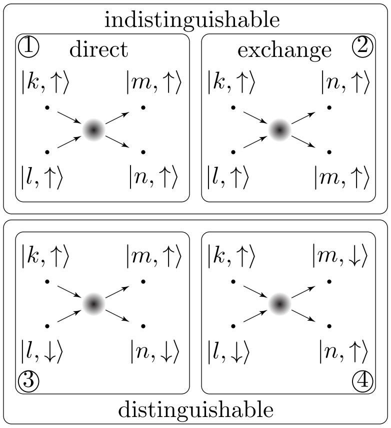


Fig. 20.11 Electron-electron scattering. Since the direct event ① is indistinguishable from the exchange event ②, their amplitudes must be added with a negative sign because they are fermions. Because of different spins of the initial states on the other hand, events ③ and ④ are distinguishable. The total matrix element squared for all four processes thus will be $|M_{\textcircled{1}} - M_{\textcircled{2}}|^2 + |M_{\textcircled{3}}|^2 + |M_{\textcircled{4}}|^2$, where $M_{...}$ denotes the matrix element for each process.

Scattering of an electron from a defect potential causes no identity crisis, because the defect and the electron are distinguishable. But for scattering of an electron with another electron, we must account for the possibility of fundamentally indistinguishable outcomes. Similarly, scattering of bosonic particles must account for their indistinguishability. The fermionic or bosonic nature of the two types of quantum particles must be accounted for in the scattering rates found from Fermi's golden rule. Consider case ① for electron-electron scattering in Fig. 20.11. The incident electrons are in states, or orbitals $|k\rangle$ and $|l\rangle$, and both have spin up $|\uparrow\rangle$. We write the composite initial states compactly as $|k,\uparrow\rangle$ and $|l,\uparrow\rangle$. The scattering event loads them into final states (or orbitals) $|m,\uparrow\rangle$ and $|n,\uparrow\rangle$. For an interaction potential W, Fermi's golden rule gives the net rate of energy-conserving scattering as

$$S(k,l \to m,n) = \frac{2\pi}{\hbar} |\langle \psi_{m,n} | W | \psi_{k,l} \rangle|^2 \delta(E_k + E_l - E_m - E_n), \quad (20.35)$$

where the amplitude of scattering is the matrix element $\langle \psi_{m,n} | W | \psi_{k,l} \rangle$. This amplitude involves the 2-particle initial wavefunction $\psi_{k,l}(r_1, r_2) = \phi_k(r_1)\phi_l(r_2)$, a product of the initial state orbital wavefunctions, prepared in a known state. The final states of events ① and ②, labeled direct and exchange in Fig. 20.11 are indistinguishable. This is captured by writing the 2-particle final wavefunction as the **antisymmetrized** product of two 1-electron orbital wavefunctions

$$
\begin{gathered}
\psi_{m,n}(r_1,r_2) = \frac{1}{\sqrt{2}}[\phi_m(r_1)\phi_n(r_2) - \phi_m(r_2)\phi_n(r_1)] \implies \\
\langle \psi_{m,n}|W|\psi_{k,l}\rangle = \frac{1}{\sqrt{2}}\langle \phi_m^\star(r_1)\phi_n^\star(r_2) - \phi_m^\star(r_2)\phi_n^\star(r_1)|W|\phi_k(r_1)\phi_l(r_2)\rangle = \\
\underbrace{\frac{1}{\sqrt{2}}\langle \phi_m^\star(r_1)\phi_n^\star(r_2)|W|\phi_k(r_1)\phi_l(r_2)\rangle}_{M_{\textcircled{1}}} - \underbrace{\frac{1}{\sqrt{2}}\langle \phi_m^\star(r_2)\phi_n^\star(r_1)|W|\phi_k(r_1)\phi_l(r_2)\rangle}_{M_{\textcircled{2}}}, \\
\text{direct integral } M_{\textcircled{1}} = \frac{1}{\sqrt{2}}\int dr_1 dr_2 \phi_m^\star(r_1)\phi_n^\star(r_2) W \phi_k(r_1)\phi_l(r_2), \\
\text{and exchange integral } M_{\textcircled{2}} = \frac{1}{\sqrt{2}}\int dr_1 dr_2 \phi_m^\star(r_2)\phi_n^\star(r_1) W \phi_k(r_1)\phi_l(r_2).
\end{gathered}
\tag{20.36}
$$

Because the processes ① and ② are indistinguishable for fermions, the matrix element that enters the Fermi's golden rule result is $(M_{\textcircled{1}} - M_{\textcircled{2}})$. The matrix elements for the distinguishable events are $M_{\textcircled{3}} = M_{\textcircled{1}}$ and $M_{\textcircled{4}} = -M_{\textcircled{2}}$. As a result, the net electron-electron scattering rate for all four events in Fig. 20.11 is

$$
\begin{aligned}
S(k,l \to m,n) = \frac{2\pi}{\hbar}(|M_{\textcircled{1}} - M_{\textcircled{2}}|^2 + |M_{\textcircled{1}}|^2 + |M_{\textcircled{2}}|^2) \times \\
f_k f_l (1-f_m)(1-f_n) \cdot \delta(E_k + E_l - E_m - E_n).
\end{aligned}
\tag{20.37}
$$

This form takes into account the probability $f_k f_l$ that the initial states are occupied, and $(1-f_m)(1-f_n)$ that the final states are empty for this scattering to occur. The reason for the form $|M_{\textcircled{1}} - M_{\textcircled{2}}|^2 + |M_{\textcircled{1}}|^2 + |M_{\textcircled{2}}|^2$ is seen in Fig. 20.11. Because ① and ② are indistinguishable, the processes interfere. Thus, their amplitudes must be subtracted first before squaring the modulus to find their intensity. Because ③ and ④ are distinguishable from each other and from processes ① and ②, their intensities should be added[9].

For boson-boson scattering, the final 2-particle state is written as $\psi_{m,n}(r_1,r_2) = \frac{1}{\sqrt{2}}[\phi_m(r_1)\phi_n(r_2) + \phi_m(r_2)\phi_n(r_1)]$, the **symmetrized** linear combination. This leads to a scattering rate

$$
\begin{aligned}
S(k,l \to m,n) = \frac{2\pi}{\hbar}(|M_{\textcircled{1}} + M_{\textcircled{2}}|^2 + |M_{\textcircled{1}}|^2 + |M_{\textcircled{2}}|^2) \times \\
f_k f_l (1+f_m)(1+f_n) \cdot \delta(E_k + E_l - E_m - E_n),
\end{aligned}
\tag{20.38}
$$

with the same expressions for the direct and exchange integrals as in Equation 20.36. Note that for bosons, occupation function terms such as f_k can be greater than unity, unlike for fermions for which $0 \leq f_k \leq 1$ because of the Pauli exclusion principle.

[9] Such scattering is encountered in semiconductors either for high electron concentrations, called **Auger recombination**, or in pair-production processes under high electric fields, called **impact ionization**. We will encounter these solid-state analogues of high-energy quantum physics in Chapter 24.

20.7 Dyson series and diagrams

In going from Equation 20.9 to 20.10, we had unceremoniously abandoned the higher order interaction terms. For most cases, this will serve us well. For the rest of the cases, here is how things can be made to work. Rewriting Equation 20.9 with the slow-turn on perturbation potential $\hat{W}_t = e^{\eta t}\hat{W}$, we get:

$$\langle n|\Psi(t)\rangle = \frac{1}{i\hbar}\int_{t_0}^{t} dt' \langle n|\hat{W}(t')|0\rangle + \frac{1}{(i\hbar)^2}\int_{t_0}^{t} dt' \langle n|\hat{W}(t')\int_{t_0}^{t'} dt''\hat{W}(t'')|0\rangle + ... \tag{20.39}$$

The *interaction representation* gives the perturbed quantum state at time t as $|\psi_t\rangle = e^{-i\frac{\hat{H}_0}{\hbar}t}|\psi(t)\rangle$, where $\hat{H}_0$ is the unperturbed Hamiltonian *operator*. This step helped us recast the time-dependent Schrödinger equation $i\hbar\frac{\partial}{\partial t}|\psi_t\rangle = (\hat{H}_0 + \hat{W}_t)|\psi_t\rangle$ to the simpler form $i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \hat{W}(t)|\psi(t)\rangle$, where $\hat{W}(t) = e^{+i\frac{\hat{H}_0}{\hbar}t}\hat{W}_t e^{-i\frac{\hat{W}_0}{\hbar}t}$ is the time-evolution operator. This equation was integrated over time to yield the Dyson series

$$\begin{aligned}|\psi(t)\rangle = \underbrace{|0\rangle}_{|\psi(t)\rangle^{(0)}} &+ \underbrace{\frac{1}{i\hbar}\int_{t_0}^{t} dt'\hat{W}(t')|0\rangle}_{|\psi(t)\rangle^{(1)}} + \underbrace{\frac{1}{(i\hbar)^2}\int_{t_0}^{t} dt'\int_{t_0}^{t'} dt''\hat{W}(t')\hat{W}(t'')|0\rangle}_{|\psi(t)\rangle^{(2)}} \\ &+ \underbrace{\frac{1}{(i\hbar)^3}\int_{t_0}^{t} dt'\int_{t_0}^{t'} dt''\int_{t_0}^{t''} dt'''\hat{W}(t')\hat{W}(t'')\hat{W}(t''')|0\rangle}_{|\psi(t)\rangle^{(3)}} + ...,\end{aligned} \tag{20.40}$$

where $|\psi(t_0)\rangle = |0\rangle$ is the initial state. Restricting the Dyson series to the first order term in $\hat{W}$ for a perturbation of the form $\hat{W}_t = e^{\eta t}W(r)$, we derived Fermi's golden rule for the transition rate $\Gamma^{(1)}_{0\to n} = \frac{2\pi}{\hbar}|\langle n|W(r)|0\rangle|^2\delta(\epsilon_0 - \epsilon_n)$. We used $\lim_{\eta\to 0^+}\frac{2\eta}{x^2+\eta^2} = 2\pi\delta(x)$ in this process. The second and third order terms in W in the Dyson series lead to a modified golden rule result

$$\begin{aligned}\Gamma_{0\to n} = \frac{2\pi}{\hbar}|\langle n|\hat{W}|0\rangle + \sum_m \frac{\langle n|\hat{W}|m\rangle\langle m|\hat{W}|0\rangle}{\epsilon_0 - \epsilon_m + i\eta\hbar} + \\ \sum_{k,l}\frac{\langle n|\hat{W}|k\rangle\langle k|\hat{W}|l\rangle\langle l|\hat{W}|0\rangle}{(\epsilon_0 - \epsilon_k + 2i\eta\hbar)(\epsilon_0 - \epsilon_l + i\eta\hbar)} + ...|^2\delta(\epsilon_0 - \epsilon_n),\end{aligned} \tag{20.41}$$

where in the end we take $\eta \to 0^+$. We identify the Green's function propagators of the form $G = \sum_m \frac{|m\rangle\langle m|}{\epsilon_0 - \epsilon_m + i\eta\hbar}$. Thus, the result to higher orders may be written (skipping the hats) in the compact form

$$\Gamma_{0\to n} = \frac{2\pi}{\hbar}|\langle n|W + WGW + \boxed{WGWGW} + ...|0\rangle|^2\delta(\epsilon_0 - \epsilon_n). \tag{20.42}$$

The boxed term corresponds to a sum of the kind shown in Fig. 20.12, where the *virtual* states of higher order terms are shown explicitly.

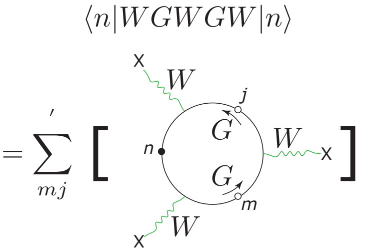


Fig. 20.12 Perturbation diagram showing a state $|n\rangle$ coming back to itself after multiple scattering events. The state n scatters into state m due to the scattering potential W. It then propagates in state m till it scatters into state j, which propagates before scattering back to state n. The amplitudes are given by the Green's function propagator G. The amplitudes of each possible choice of the intermediate states m and j must be included in the sum. The summations can sometimes be performed diagrammatically. This procedure to account for higher order interactions is also known as the "GW" approximation because of the notation.

Extensions of diagrams of this nature go under several names, the most popular of which are the space-time Feynman diagrams.

20.8 Zero-sum game: self energy

Now for something as trivial as profound: time-independent perturbation theory is a limiting case of time-dependent perturbation! Well, well – we now see that we can recover all of time-independent perturbation theory such as the Rayleigh–Schrödinger and Brillouin–Wigner theories by extending the time-evolution operator.

What is the probability that at time t after turning on a perturbation $\hat{W}_t$, we can still find the state in the *initial* state $|0\rangle$? We must find the amplitude $\langle 0|\psi_t\rangle$ and square it to get the answer. Doing so will lead to a deep connection between time-dependent and time-independent perturbation theories. In fact, it will show that time-independent perturbation theory is a *special case* of time-dependent perturbation theory – just as ″DC″ is a special case of ″AC″ with frequency $\omega \to 0$. Since $\langle 0|\psi_t\rangle = e^{-i\epsilon_0 t/\hbar}\langle 0|\psi(t)\rangle$, we can write the time-dependent Schrödinger equation as

$$i\hbar\frac{\partial}{\partial t}\ln\langle 0|\psi(t)\rangle = \langle 0|\hat{W}(t)|0\rangle + \sideset{}{'}\sum_n \langle 0|\hat{W}(t)|n\rangle\frac{\langle n|\psi(t)\rangle}{\langle 0|\psi(t)\rangle}, \tag{20.43}$$

where the prime over the sum indicates that $|n\rangle = |0\rangle$ is excluded. Let the perturbation be $\hat{W}_t = e^{\eta t}\hat{W}$. Now using the Dyson series to the 1st order in V, in the limit $\eta \to 0^+$, we get

$$i\hbar\frac{\partial}{\partial t}\ln\langle 0|\psi(t)\rangle = \langle 0|\hat{W}|0\rangle + \sideset{}{'}\sum_n \frac{|\langle n|\hat{W}|0\rangle|^2}{\epsilon_0 - \epsilon_n + i\eta\hbar}. \tag{20.44}$$

Notice that the terms on the right side look very similar to the Rayleigh–Schrödinger perturbation in energy to first and second order in $\hat{W}$. Now by integration, the required probability amplitude is

$$\boxed{\langle 0|\psi_t\rangle \sim e^{-i\frac{E_0}{\hbar}t}e^{-\frac{\Gamma}{2}t}}, \text{ where } \boxed{E_0 = \epsilon_0 + \langle 0|\hat{W}|0\rangle + P\sum_n \frac{|\langle n|\hat{W}|0\rangle|^2}{\epsilon_0 - \epsilon_n}}, \text{ and } \boxed{\Gamma = \sum_n \frac{2\pi}{\hbar}|\langle n|\hat{W}|0\rangle|^2\delta(\epsilon_0 - \epsilon_n).} \tag{20.45}$$

Here $P(...)$ is the principal part. We see that the time-independent perturbation theory may be recovered as a special case of time-dependent perturbation theory. The ″imaginary″ component of energy $\sum = i\Gamma/2$ is called the *self-energy* of state $|0\rangle$. Note that the probability is conserved at all times : i.e., $\sum|\langle n|\psi_t\rangle|^2 = 1$.

20.9 Chapter summary section

In this chapter, we have learned how to quantitatively compute various transition rates from quantum mechanics. Specifically, we learned:

- Fermi's golden rule scattering rate approximated by $\frac{1}{\tau} = \frac{2\pi}{\hbar}|W_{if}|^2\delta(E_f - E_i)$ for time-independent (static) scattering potentials. The δ-function conserves energy, and the matrix element $W_{if} = \langle f|W(r)|i\rangle$ ensures momentum conservation. The scattering matrix element for plane waves is the Fourier transform of the scattering potential, a result called the Born approximation.
- The scattering rate increases as the square of the scattering potential. If the first order scattering is prohibited by the matrix element, higher order scattering may still be possible.
- The scattering rate of electrons from time-varying scattering potentials such as light and phonon waves is $\frac{1}{\tau} = \frac{2\pi}{\hbar}|W_{if}|^2\delta(E_f - E_i \pm \hbar\omega)$ showing both absorption and emission processes.
- Transitions into a continuum of final states makes the rate $\frac{1}{\tau} \propto g(E)$ proprtional to the density of final states $g(E)$.
- Fermi's golden rule can provide understanding of conductivity of solids, and decoherence processes in qubits.

Further reading

Excellent discussion of time-dependent perturbation theory appears in the quantum mechanics textbooks to which references were provided in Chapter 3 of this book. Sakurai's **Modern Quantum Mechanics** in particular presents several physical explanations of various finer points of Fermi's golden rule, and associated rates of transitions in a rich variety of quantum systems. Kroemer's **Quantum Mechanics** presents this topic from the point of view of semiconductor physics at a level of rigor and physical insight not found in many texts. Landsberg's **Recombination in Semiconductors** has a clear exposition of the effect of many-particle statistics on transitions.

Exercises

(20.1) **Practicing Fermi's golden rule without a license**
In his textbook on Quantum Mechanics, Herbert Kroemer states: *"Practicing the golden rule without a license is a common offense, though it tends to pale in comparison to the misuse of the uncertainty relations"*. Explain the reason for this stern word of caution!

(20.2) **Exponentials of matrices**
In the derivation of Fermi's golden rule, we in-

voked the exponential of the Hamiltonian matrix. Answer the following questions using the results discussed in Chapter 6, Section 6.4:

(a) Discuss why $\exp[\beta \hat{H}_0]$, the exponential of a $N \times N$ Hamiltonian matrix $\hat{H}_0$ is also a $N \times N$ matrix. Here $\beta = 1/(k_b T)$, inverse of the thermal energy, keeps the argument of the exponential dimensionless.

(b) If the eigenvalues of the Hamiltonian matrix are $\lambda_1, \lambda_2, ..., \lambda_N$, then show that $\mathcal{Z} = \text{Trace}[\exp[\beta \hat{H}_0]] = \exp[\beta\lambda_1] + \exp[\beta\lambda_2] + ... \exp[\beta\lambda_N]$, i.e., the Trace of the exponential of the Hamiltonian is the sum of the exponentials of the individual eigenvalues. This quantity plays an important role in quantum statistical mechanics, and is called the **partition function**.

(c) Show that the determinant $\text{Det}[\exp[\beta \hat{H}_0]] = \exp[\beta\mathcal{U}]$ where $\mathcal{U} = \sum_{i=1}^{N} \lambda_i$ is the sum of the eigenvalues.

(d) In perturbation theory, the exponential of the Hamiltonian with a time-dependent perturbation $\hat{H}_0 + W(r,t)$ appears in the exponent and mixes, or "scatters" the unperturbed eigenstates. The matrix approach leads to the density matrix formalism and Green's function approaches using the Kubo formalism. Read about this approach and appreciate that the Fermi's golden rule is an approximation to this exact formulation.

(20.3) Matrix elements and selection rules

The matrix element in Fermi's golden rule is of the form $\langle f|W(r)|i\rangle = \int d^3r \psi_f^\star(r) W(r) \psi_i(r)$, where $W(r)$ is the spatial dependence of the perturbation potential, and $\psi_i(r)$ and $\psi_f(r)$ are the wavefunctions of the initial and final states.

(a) Discuss the conditions when this matrix element becomes zero by symmetry. For example, if $\psi_i(r)$ is a s-like orbital spherically symmetric around the origin, and $\psi_f(r)$ is a p_x-like orbital antisymmetric around the origin, what symmetry must the perturbation potential possess so that the matrix element is non-zero? This criteria is called the selection rule for the transition.

(b) The Dirac delta function of the Fermi's golden rule maintains energy conservation, whereas the matrix element maintains the conservation of momentum. To see why, consider a transition from the plane wave state $\psi_{k_i}(x) = e^{ik_i x}$ to the state $\psi_{k_f}(x) = e^{ik_f x}$. Show then that if $W(x) = W_0 e^{iqx}$ has a momentum component $\hbar q$, then only when $\hbar k_f = \hbar k_i + \hbar q$ is the matrix element non-zero. This strongly affects transport properties in Chapter 23 and in optical transitions in Chapter 27.

(c) Since the scattering rate is proportional to the **modulus squared** of the matrix element, argue why for example the rate is not affected by the sign of charges that are scattered.

(20.4) A sudden perturbation

Here we model a time-dependent phenomenon encountered in semiconductor devices that (strangely?) does not need the machinery of time-dependent perturbation theory and is solved using the time-independent formalism of Chapter 7. We solve the general problem first, and then apply to specific examples.

(a) Consider an electron in the n^{th} eigenstate $|n_i\rangle$ of eigenvalue $E_{n,i}$ due to the Hamiltonian $\hat{H}_i$. A sudden perturbation changes the Hamiltonian to $\hat{H}_i \to \hat{H}_i + \hat{W}$ instantaneously. Show that the probability that the electron is in the $|m_f\rangle$ eigenstate with eigenvalue E_{mf} is $|\langle m_f|n_i\rangle|^2$.

(b) Consider an electron initially in the n^{th} state of an infinitely deep 1D quantum well of width a, where the ground state is $n = 1$. The right side of the wall is suddenly moved from $x = a$ to $x = b > a$. Show that the probability that the electron is found in the m^{th} eigenstate of the new wider quantum well is $\frac{4ab^3}{\pi^2} \cdot \frac{n^2 \sin^2(\frac{\pi m a}{b})}{(a^2m^2 - b^2n^2)^2}$.

(c) Such situations are encountered in transistors or in optical devices using pn junction diodes in response to a very fast change in voltages in the gate, or across junctions. Discuss the implications of the probability of transitions calculated above for such semiconductor devices.

(20.5) Scattering from a step barrier

In Chapter 14, Exercise 14.3, we discussed the exact solution of the reflection probability $R(E)$ for an electron incident on a heterostructure potential barrier. An approximation solution can also be obtained by Fermi's golden rule.

(a) Make a sketch of the reflection probability of Equation 14.54 as a function of the energy E of the electron, using ΔE_c and a as paramters. What is the limit when $E >> \Delta E_c$?

(b) Show now using Fermi's golden rule that the net rate of reflection is the same as the $E >> \Delta E_c$ approximation of the exact relation of part (a). Follow though the steps since each illustrates the scope and the limitations of the Fermi's golden rule approach.

(20.6) **Stuck in a Dirac delta well: light to the rescue!**

An electron is stuck in the single bound state of a 1D Dirac delta potential well, at an energy E_δ below zero. A light beam creates an electric potential $W(x) = F_0 x \cos(\omega_0 t)$ for the electron.

(a) Show that the rate at which an electron is ejected from the well is given by $1/\tau = \frac{2\hbar F_0^2 |E_\delta|^{\frac{3}{2}} \sqrt{\hbar\omega_0 - |E_\delta|}}{m_e (\hbar\omega_0)^4}$.

Solution: A 1D Dirac delta well famously has a single bound state (see Equation 5.87) at an energy $|E_\delta| = \frac{m_e S^2}{2\hbar^2} = \frac{\hbar^2\kappa^2}{2m_e}$ below zero, and an exponentially decaying eigenfunction $\psi_\delta(x) = \sqrt{\kappa} e^{-\kappa|x|}$. Let us approximate the final state to be a free plane wave blissfully unaware of the presence of the well with eigenfunction $\psi_f(x) = \frac{1}{\sqrt{L}} e^{ikx}$ and energy $E_f = \frac{\hbar^2 k^2}{2m_e}$. Here L is an artificial large length inside which the well is located. Then the matrix element is $\langle f|W|i\rangle = \int_{-\infty}^{+\infty} dx \cdot \psi_f^\star(x) W(x) \psi_\delta(x) = -\frac{2ik\kappa^{\frac{3}{2}} F_0}{\sqrt{L}(k^2+\kappa^2)^2}$. Calculate and confirm this, using the fact that in $W(x) = \frac{F_0 x}{2}(e^{i\omega_0 t} + e^{-i\omega_0 t})$, only the absorption term can remove an electron from the bound state into the continuum; there is no emission. From Fermi's golden rule, the scattering rate then is $\frac{1}{\tau_{i\to f}} = \frac{2\pi}{\hbar} |\langle f|W|i\rangle|^2 \delta(E_f - (E_i + \hbar\omega_0))$, which when summed over all final k-states using $\sum_k (...) = \int_{-\infty}^{+\infty} \frac{dk}{\frac{2\pi}{L}} (...)$ gives $\frac{1}{\tau} = \frac{2\pi}{\hbar} \cdot 4F_0^2 \kappa^3 \int_{-\infty}^{+\infty} dk \cdot \frac{k^2}{(k^2+\kappa^2)^4} \delta(\frac{\hbar^2 k^2}{2m_e} - (\hbar\omega_0 - |E_\delta|)) = \frac{2\hbar F_0^2 |E_\delta|^{\frac{3}{2}} \sqrt{\hbar\omega_0 - |E_\delta|}}{m_e (\hbar\omega_0)^4}$.

(b) Show that the maximum rate of electron ejection for a given photon energy is $\frac{3\sqrt{3}\hbar F_0^2}{8 m_e (\hbar\omega_0)^2}$, whereas for a given bound state energy it is $\frac{343\sqrt{7}\hbar F_0^2}{2048 m_e |E_\delta|^2}$.

(c) Electrons trapped in deep level states in the bandgap of a 1D semiconductor quantum wire can be optically excited into the conduction band in this fashion. Consider $N_t \sim 10^5$/cm deep traps in a quantum wire, $|E_\delta| = 1$ eV below the conduction band edge in a wide bandgap semiconductor quantum wire. Find and plot the generation rate of electrons in G in (cm·s)$^{-1}$ as a function of the photon energy $\hbar\omega_0$ and electric field amplitude F_0/q.

(20.7) **Scattering rate due to a large number of randomly distributed scattering sites**

Instead of one scatterer of potential $W_0(r)$ centered at the origin, how would the net scattering rate depend if there were N scatterers distributed randomly at $r = R_i$ for $i = 1$ to N, with a net scattering potential $W(r) = W_0(r - R_1) + W_0(r - R_2) + ... + W_0(r - R_N)$? Show that $\frac{1}{\tau}$ is proportional not to N^2 but to N. This is encountered in ionized impurity scattering, and is discussed in Chapter 23, and leads to the experimental result that the ionized-impurity limited electron mobility in a semiconductor reduces as $\mu_{imp} \propto 1/N_{imp}$.

(20.8) **Optical transition rate of a hydrogen atom**

As a model of transition from bound to free states, consider an electron in the ground state of a hydrogen atom in its initial state. The energy is $E_{1s} = -Ry$ where Ry is the Rydberg, and the wavefunction is $\psi_{1s}(r) = \frac{1}{\sqrt{\pi a_B^{\frac{3}{2}}}} \exp[-\frac{r}{a_B}]$, where a_B is the Bohr radius. Assume that a light beam creates the perturbation potential $W(z,t) = qFz\cos\omega t$.

(a) Find the matrix element for the transition from the ground state to a free-electron final state with energy $\frac{\hbar^2 k_f^2}{2m_e}$ and wavefunction $\psi_f(r) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}$.

(b) Sketch the shape of the squared matrix element. For wavelengths of the final state is the matrix element maximized?

(c) Summing over all possible final states using the Fermi-golden rule, calculate and estimate numerically the ionization rate of the hydrogen atom for $F = 1$ kV/m.

(20.9) **The Kubo–Greenwood conductivity formula**

In this problem, we show that the Boltzmann result for metallic conductivity $\sigma = nq^2\tau/m_e$ can be obtained from the Kubo–Greenwood formula that is derived from Fermi's golden rule. It was shown in the chapter why the low-temperature DC conductivity from the Kubo–Greenwood formula is $\sigma(0) \approx \frac{2\pi q^2 \hbar^3 V}{m_e^2} |D|_{av}^2 g(E_F)^2 \approx nq^2\tau/m_e$ up to a factor of unity.

(a) Show that the averaged matrix element square $|D|_{av}^2 \approx \frac{\pi\lambda_{mfp}}{3V}$, where the mean free path $\lambda_{mfp} = v_F\tau$ is the product of the Fermi velocity v_F and a scattering time τ.

Solution: The matrix element in Fermi's golden rule is $D = \int d^3x \cdot \psi_{E+\hbar\omega} \frac{\partial}{\partial x} \psi_E$. Consider the initial and final states to be plane waves $\psi_E(r) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}}$ and $\psi_E(r) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}'\cdot\mathbf{r}}$. The matrix element is then $D = \frac{i\mathbf{k}}{V} \int d^3r e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}$, where the integral is over the entire volume V. If there is a scattering event every τ second for electrons at the Fermi surface moving with velocity v_F, the mean free path is $\lambda_{mfp} \sim v_F \tau$. As a result of scattering, the wavevector $\mathbf{k}$ is randomized over a propagation length λ_{mfp}, and volume $v = \frac{4}{3}\pi\lambda_{mfp}^3$. There are $N = V/v$ such volumes in the crystal. The integral $\int_V d^3r e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}} = \int_v d^3r e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r_1}} + \int_v d^3r e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r_2}} ... + \int_v d^3r e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r_N}}$ is thus a sum of N terms, with uncorrelated, randomized phases because of scattering. If scattering occurs reasonably often, the volume v is small, $e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r_1}} \approx e^{i\phi_1}$ is constant within volume v, and $\int_v d^3r e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r_1}} \approx v e^{i\phi_1}$. Then, $D = i\mathbf{k} \cdot \frac{v}{V}(e^{i\phi_1} + e^{i\phi_2} + ... + e^{i\phi_N})$, and $|D|^2 = k^2(\frac{v}{V})^2 N = \frac{v}{V}k^2$, where because of the random phases and large N, the cross terms cancel. This sort of summation is characteristic of a random walk or diffusive process. We must average this $|D|^2$ over a volume v and not V. Assuming the scattering to be nearly elastic, $(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r} \approx 2k\sin(\frac{\theta}{2})\lambda_{mfp} \approx k\theta\lambda_{mfp} \leq 1$, meaning the small angular deviation is $\theta \leq 1/(k\lambda_{mfp})$. Averaging over the small angle cone yields $|D|^2_{av} = \frac{v}{V} k^2 \frac{\int_0^{1/k\lambda_{mfp}} \sin(\theta) d\theta}{\int_0^{2\pi} \sin(\theta) d\theta} \sim \frac{v}{V} \cdot \frac{1}{4\lambda_{mfp}^2} = \frac{\pi\lambda_{mfp}}{3V} = \frac{\pi v_F \tau}{3V}$.

(b) Using the value of $|D|^2_{av}$ from part (a) and the metallic density of states $g(E_F)$, prove that the DC conductivity is indeed $\sigma(0) = nq^2\tau/m_e$ up to a factor of unity.

(c) D. J. Thouless took a different approach (see *The Philosophical Magazine*, volume 32, Issue 2, pages 877-879, 1975) to derive the entire AC response from the Kubo–Greenwood formula as

$$\sigma(\omega) = \sigma(0) \cdot \frac{1}{1 + \frac{\frac{1}{2} m_e \omega^2 \lambda_{mfp}^2}{\frac{\hbar^2 k_F^2}{2m_e}}} = \frac{\sigma(0)}{1 + \omega^2\tau^2}. \quad (20.46)$$

Identify τ in the AC damping of conductivity with the scattering time. Read the paper and note the rather interesting approach taken in it.

(20.10) **Rabi oscillations**

In this Exercise, you solve the 2-level quantum system problem exactly to recover the results of Rabi oscillations.

(a) Prove that the solution to the matrix Equation 20.33 is given by Equations 20.34.

(b) Make numerical plots of the solutions as shown in Fig. 20.8 for $\hbar\omega_{21} = 1.0$ eV and for values of $\gamma = 0.01, 0.1,$ and 1.0 eV.

(c) Do some research on the range of physical problems for which the exact 2-level solution has enabled practical applications, and write a short report on its importance in the past and in the future.

No Turning Back: The Boltzmann Transport Equation

21

21.1 Micro vs. macro 481
21.2 The Liouville theorem 483
21.3 Boltzmann transport equation 487
21.4 H-theorem and entropy 490
21.5 Equilibrium distribution 493
21.6 The RTA: time to relax! 499
21.7 One formula to rule them all 501
21.8 Electrical conductivity 506
21.9 Thermoelectric properties 511
21.10 Onsager relations 517
21.11 Conservation laws 518
21.12 Berry curvature correction 520
21.13 Limitations of the BTE 521
21.14 Chapter summary section 523
Further reading 523
Exercises 524

We have achieved a lot in the past chapters by investigating ballistic transport and paying only scant attention to electron scattering processes in semiconductors. The time has come to discuss scattering and its deep connection with the concepts of equilibrium, irreversibility, and entropy. The Boltzmann transport equation (**BTE**) is the framework to introduce the effect of scattering processes in solids on a common platform that bakes in the quantum mechanics of electrons and phonons with thermodynamics and electromagnetism. A unified picture to understand the flow of energy in solids carried by charge and heat and their connection will emerge. We will see how the BTE introduces and enforces irreversibility and entropy into reversible quantum mechanical transport processes. In this chapter, we learn:

- Why do we need the BTE to understand semiconductors?
- How does the BTE help calculate transport properties?
- What are the power and limitations of the BTE?

21.1 Micro vs. macro

The current I we measure flowing in a semiconductor is a macroscopic value at the contacts. The macroscopic value is what ends up at the boundaries from the sum total of all microscopic currents flowing inside the bulk. The microscopic processes respond to external controls such as voltage bias at contacts, electric field effect through insulators, magnetic fields, temperature gradients, and photonic illumination. In this sense, the current I is a quantity similar to the pressure P of a gas, whose microscopic origin is in the collisions of molecules with the wall of a container and amongst themselves. To connect the microscopic processes of electrons to the measurable macroscopic current, we will be guided by the same kinetic theory that explains the behavior of gases.

Let us start with a simple question: an electron is subjected to a constant external electric field F. What are its transport properties? Consider the *classical* mechanics of an electron starting from rest $v(t =$

0) = 0 at $x(t=0)=0$ in vacuum. The answer is precise: by Newton's law, $-qF = m_e \frac{dv(t)}{dt}$ dictates that the electron velocity $v(t) = -\frac{q}{m_e}Ft$ increases in magnitude linearly with time t, i.e., the electron undergoes uniform acceleration. Its location and momentum $[x(t), m_e v(t)]$ follow a sharply defined, deterministic path in the $[x,p]$ space with time.

How does the electron dynamics change in quantum mechanics? The eigenfunctions and eigenvalues allowed for a free–electron in vacuum in the absence of an electric field are $\psi(r) = e^{i\mathbf{k}\cdot\mathbf{r}}/\sqrt{V}$ and $E(k) = \hbar^2 k^2/2m_e$. This set does not allow the initial conditions $[x(t=0) = 0, v(t=0) = 0]$ due to the uncertainty principle rooted in the wave-particle duality. However, since the allowed eigenfunctions form a complete set, quasi-localized states can be constructed by the linear combinations $\psi(r,t) = \sum_k c_k e^{i(\mathbf{k}\cdot\mathbf{r} - \frac{\hbar k^2}{2m_e}t)}$, which are **wavepackets**, such that their center coordinates are well defined. The electric field will then change the center $\mathbf{k}$ and $\mathbf{r}$ of the electron wavepacket with time. The quantum state will be composed of a *range* of $[\mathbf{r},\mathbf{k}]$ values. In the BTE language, the $[\mathbf{r},\mathbf{k}]$ space is called the **phase space**, illustrated in Fig. 21.1 for 1D. Because there is only one electron, the quantum mechanical wavepacket solution $\psi(r,t)$ must satisfy[1] for the general 3D case $\int d^3r \cdot \frac{d^3k}{(\frac{2\pi}{L})^3} \cdot |\psi(r,t)|^2 = 1$ at all times t. Because there is just one electron, there is no Pauli blocking due to the exclusion principle. The trajectory in the phase space is diffuse compared to the sharp classical path due to the uncertainty principle $\Delta x \cdot \Delta k \geq \frac{1}{2}$. The "center" of the wavepacket tracks the path given by $\mathbf{F} = \hbar d\mathbf{k}/dt$.

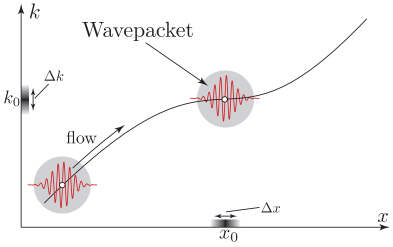


Fig. 21.1 Illustration of the motion of a wavepacket in the phase–space.

Now consider the single electron residing in a crystal. The periodic crystal potential $V(r+a) = V(r)$ rearranges the electron energy eigenvalues into bands $E_n(k)$ with corresponding Bloch eigenfunctions $\psi_k(r) = e^{ikr}u_k(r)$. Fig. 21.2 indicates the single electron in a band. In Chapter 9 we learned that the dynamics of a **Bloch wavepacket** $\psi(r,t) = \sum_k c_k \psi_k(r) e^{-i\frac{E_n(k)}{\hbar}t}$ in the phase space is given by[2]

$$\boxed{\mathbf{v}(\mathbf{k}) = \frac{d\mathbf{r}}{dt} = \underbrace{\frac{1}{\hbar}\nabla_k E_n(k)}_{\mathbf{v}_g(\mathbf{k})} - \frac{\mathbf{F}}{\hbar} \times \mathbf{\Omega_k}} \;\&\; \boxed{\frac{d(\hbar\mathbf{k})}{dt} = (-q)\underbrace{[\mathbf{E} + \mathbf{v}_g(\mathbf{k}) \times \mathbf{B}]}_{\text{Lorentz Force}},} \tag{21.1}$$

where $[\mathbf{r},\mathbf{k}]$ is the center of the Bloch wavepacket in the phase space, indicated by $[x_0, k_0]$ in Fig. 21.1, and $\mathbf{\Omega_k}$ is the Berry curvature of the band at wavevector $\mathbf{k}$. For the constant electric field $\mathbf{E}$, with $\mathbf{B} = 0$ and in a crystal for which $\mathbf{\Omega_k} = 0$, the center of the wavepacket will change linearly with time as $\mathbf{k}(t) = \mathbf{k}(0) - \frac{q\mathbf{E}}{\hbar}t$. If $\mathbf{k}(0) = 0$, the electron starts at the bottom of the band with a zero group velocity. From Fig. 21.2 it can then be seen that the electron group velocity given by the slope $v_g(k) = \frac{1}{\hbar}\frac{dE_n(k)}{dk}$ will initially increase, reach a maximum near the middle of the band, and then decrease to zero as the state reaches the top of the band, or the Brillouin zone edge at $k = \pm\frac{\pi}{a}$. It will re-enter the BZ with an opposite velocity direction. The electron will

[1] Solving the time-dependent Schrödinger equation $i\hbar\frac{\partial\psi}{\partial t} = [-\frac{\hbar^2}{2m_e}\nabla^2 + V(r)]\psi$, where $V(r) = -q\mathbf{F}\cdot\mathbf{r}$ yields the dynamical solution $\psi(r,t)$. The details of the solution reside in the coefficients c_k of the wavepacket.

[2] This is valid so long as the electron stays within the same band and does not transition to another. Such situations are described in Section 21.13 at the end of this chapter. The discussion in this chapter is restricted to relatively weak electric and magnetic fields and no optical excitation such that electron transport occurs in a single band.

then perform oscillations in the **k**-space and in energy $E_n(k(t))$, which translates also to oscillations in the real space. However, these oscillations, predicted by Bloch as a remarkable consequence of quantum mechanics of electrons in periodic crystals (see Problems 8.8 and 9.5), are rarely observed in practice.

Why are Bloch oscillations not commonplace? The first reason is because scattering processes in crystals throw the electron off the ballistic trajectory[3]. As shown in the bandstructure of Fig. 21.2, an electron can scatter from state k to k', at a rate $S(k \to k')$, preventing oscillations. **Such scattering is only possible by processes that break the periodicity of the crystal potential, such as defects and phonons**[4]. The effective mass equation *already includes* the crystal potentials of bulk semiconductors *and* perfect semiconductor heterostructures.

Even if we removed all scattering of the first kind, the second reason Bloch oscillations are not common in real solids is the presence of not just one, but a **large number** of electrons in the band. The transport mechanism and scattering are significantly affected by the **Pauli exclusion principle**, and by **electron-electron scattering**. In typical metals, the density is of the order $n \sim 10^{23}/\text{cm}^3$, which fills a significant part of the Brillouin zone, and in a semiconductor, $n \sim 10^{15} - 10^{20}/\text{cm}^3$ fill a fraction. These large numbers make it difficult to track the $[\mathbf{r}, \mathbf{k}]$ of *each* electron in time using Equations 21.1. Furthermore, to find the macroscopic properties such as currents, we sum over all **k**-states and throw away the details, making the exercise unnecessary as well! We therefore change our point of view to a far more powerful method for connecting the large numbers in the microscopic picture to the macroscopic transport coefficients such as electrical or thermal conductivity.

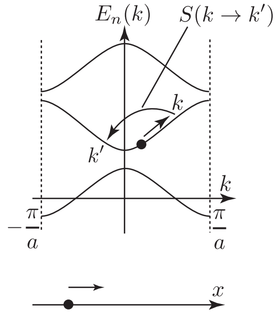


Fig. 21.2 Scattering of an electron from a state k to k' at a rate $S(k \to k')$ given by Fermi's golden rule.

21.2 The Liouville theorem

An electron wavepacket is also called a **quasiparticle** to remind us that by itself it does not have a well defined $[x, k]$, but the *center* of the wavepacket does[5]. Quantum mechanics is responsible for how the wavepacket came about in the first place: through a linear superposition of Bloch states labeled by k, which gave the wavepacket a finite spread in x. **We will refer to an electron wavepacket simply as an "electron" in a slight abuse of language.**

Instead of tracking the $[x, k]$ of each electron in the semiconductor band, we will end up tracking the *probability density* $f(x, k, t)$ of finding a collection of electrons in the neighborhood of $[x, k]$ at time t. But let's make the transition from knowing everything about every electron, to a restricted collective knowledge in steps. The individual $[\mathbf{r}_i, \mathbf{k}_i]$ of each of the $i = 1, 2, ..., N$ electrons define a $6N$-dimensional phase space called the **Γ-space**. Then, the state of the N-electron system at any given time is given by a **point** in this $6N$-dimensional Γ-space. Several distinct points in the Γ-space can correspond to the same macroscopic property of the N-electrons, such as their total energy, or the macroscopic current carried by all of them. Distinct microscopic states (or points) that lead to the same macroscopic property are called **ensembles**, and are represented by a set of points in the Γ-space. This enormous $6N$-dimensional phase space, and the concept of ensembles was introduced by Gibbs. The **Liouville theorem** discussed in this section tracks the motion of $\rho(x,k,t) = \rho(\mathbf{r}_1, \mathbf{r}_2, ...\mathbf{r}_N, \mathbf{k}_1, \mathbf{k}_2, ...\mathbf{k}_N, t)$, the many-electron distribution function, in the Γ-space when subjected to the laws of classical dynamics[6].

---

[3] The time the wavepacket needs to traverse the entire BZ is $t = h/qFa \sim 100$ ps for a small field $F \sim 1$ kV/cm. Typical scattering times are sub-ps. See Exercise 21.1.

[4] It is important to distinguish which potentials scatter electrons, and which do not. In Chapter 14 we learned the *effective mass approximation* with wavepacket eigenfunctions. In a semiconductor heterostructure, the band offsets are already included in the potential. For electrons confined in a heterostructure quantum well with perfect interfaces, in the effective mass approximation a heterojunction band offset is *not a scattering potential*. Impurities such as dopants or interface roughness will then scatter electrons, not the band offsets.

[5] In this section, we interchangeably use x and k for the vectors $\mathbf{r}$ and $\mathbf{k}$ to avoid excessive symbols, but it should be clear from the context when vectors are appropriate.

The number of points representing the electrons in a small neighborhood of $[x,k]$ in the $6N$-dimensional Γ-space at time t is $\rho(x,k,t) \cdot dx \cdot dk$. The shaded area in Fig. 21.3 is $dx \cdot dk$. When these electrons are subjected to a force $F = -dV/dx$ due to a potential V, each phase point moves to a new $[x',k']$ an instant dt later. What is the new area $dx' \cdot dk'$ at time $t + dt$, i.e., how does the **phase–space volume** change in time? Note that this is indeed a new point of view, where we track a *collective* distribution function, rather than an individual electron.

The answer to the question posed above was first found by Joseph Liouville (Fig. 21.4), and is called the Liouville theorem of classical mechanics. If there is no scattering, the rather remarkable fact is that the phase–space volume is conserved, meaning $dx' \cdot dk' = dx \cdot dk$, the area of the shaded regions in Fig. 21.3 remains the same, though it may deform: **phase space flows as an incompressible fluid**. To see why, let us track the phase space point $[x_\alpha, k_\alpha]$ (where α is a coordinate index) in time. Writing the classical momentum as $p_\alpha = mv_\alpha = \hbar k_\alpha$, $dx_\alpha/dt = \dot{x}_\alpha$ and $dp_\alpha/dt = \dot{p}_\alpha$, the classical Hamiltonian is the energy $H = p^2/2m_e + V$, and retaining terms first order in dt, we have

$$x'_\alpha = x_\alpha + \dot{x}_\alpha dt + \mathcal{O}(dt^2), \text{ and } dx'_\alpha = dx_\alpha + \frac{\partial \dot{x}_\alpha}{\partial x_\alpha} dx_\alpha dt + \mathcal{O}(dt^2)$$

$$p'_\alpha = p_\alpha + \dot{p}_\alpha dt + \mathcal{O}(dt^2), \text{ and } dp'_\alpha = dp_\alpha + \frac{\partial \dot{p}_\alpha}{\partial p_\alpha} dp_\alpha dt + \mathcal{O}(dt^2)$$

$$\implies dx'_\alpha \cdot dp'_\alpha = dx_\alpha \cdot dp_\alpha \cdot (1 + dt(\underbrace{\frac{\partial \dot{x}_\alpha}{\partial x_\alpha} + \frac{\partial \dot{p}_\alpha}{\partial p_\alpha}}_{\Delta}) + \mathcal{O}(dt^2))$$

$$\boxed{\dot{x}_\alpha = +\frac{\partial H}{\partial p_\alpha} \text{ and } \dot{p}_\alpha = -\frac{\partial H}{\partial x_\alpha}} \implies \frac{\partial \dot{x}_\alpha}{\partial x_\alpha} = +\frac{\partial^2 H}{\partial x_\alpha \partial p_\alpha} \;\&\; \frac{\partial \dot{p}_\alpha}{\partial p_\alpha} = -\frac{\partial^2 H}{\partial p_\alpha \partial x_\alpha}$$

$$\implies \Delta = 0 \implies \boxed{dx'_\alpha \cdot dp'_\alpha = dx_\alpha \cdot dp_\alpha,} \tag{21.2}$$

which proves that the phase–space area is preserved in the dynamics. Note that $\Delta = 0$ is ensured by Hamilton's equations of classical mechanics shown in the first box, which is simply the velocity $\dot{x}_\alpha = +\frac{\partial H}{\partial p_\alpha} = \frac{p_\alpha}{m_e}$ and Newton's law $\dot{p}_\alpha = -\frac{\partial H}{\partial x_\alpha} = -\frac{\partial V}{\partial x_\alpha}$.

Because of the Liouville theorem, $\rho(x',k',t+dt) = \rho(x,k,t)$, which means that in the absence of scattering events, the state that was at $[x,p]$ at time t has moved to $[x',p']$ at time $t + dt$ in a streaming motion.

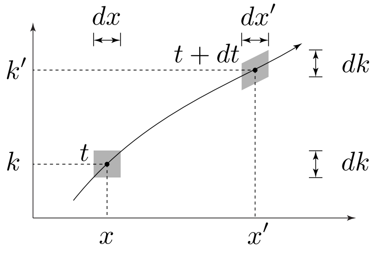


Fig. 21.3 The Liouville theorem states that the shaded area of the phase space of occupied states is preserved when the dynamics or motion of particles occupying them follow the rules of classical mechanics. The shaded area can deform due to the motion, but will preserve the area. In the absence of scattering, the trajectories of phase points in the Γ-space do not intersect.

Fig. 21.4 Joseph Liouville, French mathematician who recognized that time-evolution in systems governed by Hamiltonian equations preserves the area of phase–space.

---

[6] The full-blown quantum version tracks the density matrix $\hat{\rho}(x,k,t)$ which retains the phase relationships between the electron wavefunctions. The diagonal matrix elements of the density matrix are $\rho(x,k,t)$, the semi-classical probability density of the Liouville equation. For a majority of semiconductor transport phenomena, the density matrix is an overkill, and so is the Liouville density, so we transition in steps from $\rho(x,k,t)$ in the Γ-space to $f(x,k,t)$ in the micro (or μ-) space, where the BTE is defined.

Taylor expanding ρ with x'_α, p'_α, and dt, we can write

$$\rho(x + \dot{x} \cdot dt, p + \dot{p} \cdot dt, t + dt) = \rho(x, p, t) + \frac{\partial \rho}{\partial x} dx + \frac{\partial \rho}{\partial p} dp + \frac{\partial \rho}{\partial t} dt$$

$$\frac{d\rho}{dt} = \frac{\rho(x', k', t + dt) - \rho(x, k, t)}{(t + dt) - t} = 0 \implies \frac{\partial \rho}{\partial t} + \dot{x} \frac{\partial \rho}{\partial x} + \dot{p} \frac{\partial \rho}{\partial p} = 0$$

$$\implies \boxed{\frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{r}} \rho + \mathbf{F} \cdot \nabla_{\mathbf{p}} \rho = 0,} \text{ or equivalently, } \boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0,} \tag{21.3}$$

where the boxed version is generalized to any dimension with $\mathbf{r} = (..., x_\alpha, ...)$, $\mathbf{v} = (..., \dot{x}_\alpha, ...)$, and $\mathbf{p} = (..., p_\alpha, ...)$, and $\mathbf{J} = \rho \mathbf{v}_{tot}$ is the current density[7]. In this form, we recognize the Liouville equation as a continuity equation, which is the consequence of particle number conservation. An interesting way to write the Liouville theorem is

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \sum_\alpha \left(\frac{\partial \rho}{\partial p_\alpha} \cdot \dot{p}_\alpha + \frac{\partial \rho}{\partial x_\alpha} \cdot \dot{x}_\alpha\right) = 0$$

$$\frac{\partial \rho}{\partial t} = -\sum_\alpha \left(\frac{\partial \rho}{\partial x_\alpha} \cdot \frac{\partial H}{\partial p_\alpha} - \frac{\partial \rho}{\partial p_\alpha} \cdot \frac{\partial H}{\partial x_\alpha}\right) = -\{\rho, H\}$$

$$\implies \boxed{\frac{\partial \rho}{\partial t} + \{\rho, H\} = 0,} \tag{21.4}$$

where the $\mathbf{v} \cdot \nabla_{\mathbf{r}} \rho + \mathbf{F} \cdot \nabla_{\mathbf{p}} \rho = \nabla \cdot \mathbf{J}$ term of Equation 21.3 is identified as $\{\rho, H\}$, the **Poisson–bracket** of the distribution function ρ with the Hamiltonian H. This formulation carries over nicely to quantum mechanics[8]. The expectation value of any macroscopic physical property $g(x, p)$ (such as current, entropy, or energy) defined as a function of the points in the phase–space, is $\langle g \rangle = \langle g\rho \rangle$ for the distribution $\rho(x, p, t)$. The time evolution of its expectation value is then given by $\frac{d\langle g \rangle}{dt} = \langle \{g, H\} \rangle$. From Equation 21.4, an **equilibrium distribution function ρ_0 must be independent of time, and therefore must have** $\{\rho_0, H\} = 0$. If the equilibrium distribution function is an explicit function of the Hamiltonian or the total energy itself $\rho_0 = \rho_0(H(x, p))$, then $\{\rho_0, H\} = 0$ is a mathematical property of the Poisson bracket. Since $\rho(x, k, t)$ is the probability density, summing or integrating this function over the entire phase space gives the total number of particles at that time: $\int dxdk\rho(x, k, t) = N(t)$. If the total number of particles in the entire phase space does not change with time[9], then $\int dxdk\rho(x, k, t) = \int dxdk\rho_0(x, k) = N$.

Thus the Liouville equation, which is still exact, tracks the relative locations and momenta of the N particle ensembles in the Γ-space via the many-particle distribution function $\rho(\mathbf{r}_1, \mathbf{r}_2, ...\mathbf{r}_N, \mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_N, t)$. Let us now allow for scattering between particles, and ask how may we track the properties of a single particle, say the one with subscript 1? The single-particle distribution function for particle 1 is obtained by integrating the distribution function over the coordinates and momenta

[7] The current in this compact form is defined with $\mathbf{v}_{tot} = (\dot{\mathbf{r}}_1, \dot{\mathbf{r}}_2, ...\dot{\mathbf{r}}_N, \dot{\mathbf{p}}_1, \dot{\mathbf{p}}_2, ...\dot{\mathbf{p}}_N)$ a $6N$-dimensional vector.

[8] The **Poisson bracket** for any two functions A and B in the phase space $[x_\alpha, p_\alpha]$ in classical mechanics is defined as $\{A, B\} = \sum_\alpha (\frac{\partial A}{\partial x_\alpha} \cdot \frac{\partial B}{\partial p_\alpha} - \frac{\partial A}{\partial p_\alpha} \cdot \frac{\partial B}{\partial x_\alpha})$. Note that although A and B are *not* operators, $\{B, A\} = -\{A, B\}$. The quantum mechanical generalization of the Liouville theorem is the von Neumann formula for the density matrix, $\frac{\partial \hat{\rho}}{\partial t} = \frac{i}{\hbar}[\hat{\rho}, \hat{H}]$, where $[.,.]$ is the operator commutator.

[9] We will see soon that the equilibrium single-particle distribution function for electrons is the Fermi–Dirac function $f_0 = 1/(\exp[(E(\mathbf{k}, \mathbf{r}) - \mu)/k_b T] + 1)$, where $E(\mathbf{k}, \mathbf{r}) = H$ are indeed the energy eigenvalues of the Hamiltonian. Furthermore, **any physical conserved quantity** whose expectation value over the distribution does not change with time (called a **constant of motion**), is required to satisfy the relation $\{g, H\} = 0$ because of the relation $\frac{d\langle g \rangle}{dt} = \langle \{g, H\} \rangle$. This is the origin of the laws of conservation in physics.

of all $(N-1)$ other particles except particle 1:

$$f^{(1)}(\mathbf{r}_1, \mathbf{p}_1, t) = \frac{N!}{(N-1)!} \times$$
$$\int d\mathbf{r}_2 d\mathbf{r}_3 ... d\mathbf{r}_N d\mathbf{p}_2 d\mathbf{p}_3 ... d\mathbf{p}_N \cdot \rho(\mathbf{r}_1, ...\mathbf{r}_N, \mathbf{p}_1, ...\mathbf{r}_N, t). \tag{21.5}$$

The factorials in front are necessary because the N particles are assumed identical, and indistinguishable. If we account for inter-particle scattering between particle 1 and 2, the single particle occupation function $f^{(1)}(\mathbf{r}_1, \mathbf{p}_1, t)$ will depend on the two-particle occupation function

$$f^{(2)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{p}_1, \mathbf{p}_2, t) = \frac{N!}{(N-2)!} \times$$
$$\int d\mathbf{r}_3 d\mathbf{r}_4 ... d\mathbf{r}_N d\mathbf{p}_3 d\mathbf{p}_4 ... d\mathbf{p}_N \cdot \rho(\mathbf{r}_1, ...\mathbf{r}_N, \mathbf{p}_1, ...\mathbf{r}_N, t), \tag{21.6}$$

which in turn will depend on $f^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, t)$, the 3-particle distribution function, and so on. In the presence of collisions, the collisionless Liouville equation of Equation 21.4 evolves into a hierarchy of N coupled equations[10]. This is because the interparticle scattering potentials "entangle", or correlate the single-particle distribution functions. For example, the scattering potential $V(\mathbf{r}_1, \mathbf{r}_2)$ makes the 2-particle distribution function

$$f^{(2)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{p}_1, \mathbf{p}_2, t) = f^{(1)}(\mathbf{r}_1, \mathbf{p}_1, t) \cdot f^{(1)}(\mathbf{r}_2, \mathbf{p}_2, t) + g(\mathbf{r}_1, \mathbf{r}_2, \mathbf{p}_1, \mathbf{p}_2, t), \tag{21.7}$$

where the first term on the right is a product of single-particle distributions, but the second term does not factorize into functions for the two particles. Such correlations make the hierarchy of equations unsolvable in the presence of scattering[11].

Boltzmann made the following bold assumption: if the inter-particle correlations are weak, the corresponding factor $g(\mathbf{r}_1, \mathbf{r}_2, \mathbf{p}_1, \mathbf{p}_2, t) \approx 0$ in Equation 21.7, the N-particle distribution function factorizes into a product of single-particle functions:

$$f^{(N)}(\mathbf{r}_1, ...\mathbf{r}_N, \mathbf{p}_1, ...\mathbf{p}_N, t) = f^{(1)}(\mathbf{r}_1, \mathbf{p}_1, t) \cdot f^{(1)}(\mathbf{r}_2, \mathbf{p}_2, t) \cdot ... \cdot f^{(1)}(\mathbf{r}_N, \mathbf{p}_N, t), \tag{21.8}$$

and the hierarchy of Liouville equations truncate into a single equation, in a much lower dimensional phase space[12]. The single-particle occupation function $f^{(1)}(\mathbf{r}_1, \mathbf{p}_1, t)$ is defined in a 6-dimensional $[x, k]$ space. The N particles then occupy a **set of N independent points** in this 6-dimensional space, called the μ-space, or the **microspace**. Note the enormous conceptual simplification in going to the μ-space. We drop the superscript $f^{(1)}$ of the single-particle distribution function and denote it by $f(x, k, t)$. The Boltzmann transport equation lives in this reduced 6-D microspace. The distribution function of the BTE is thus the single-particle approximation, to which we turn our attention.

[10] This set of coupled equation is known as the BBGKY hierarchy after Bogoliubov, Born, Green, Kirkwood and Yvon.

[11] When the mean free path between scattering events is smaller than the effective size of a particle, the correlations between particles are strong and cannot be neglected. Then one must resort to molecular dynamics solutions of the Liouville equation, or use alternate formalisms, such as Einstein's stochastic model for Brownian motion. The effective size for electrons in typical semiconductors is the de Broglie wavelength. When the electron mean free path is much longer than its de Broglie wavelength, the correlations are indeed weak and the BTE is an excellent approximation.

[12] The historical name of this crucial approximation is *stosszahlansatz*, popularly referred to as the assumption of **molecular chaos**. It means that a scattering event *erases* the memory of the phase space points the particles occupied prior to scattering. The quantum analogy is when a many-particle wavefunction is written as the product of individual wavefunctions.

21.3 Boltzmann transport equation

In 1872, Boltzmann (Fig. 21.5) introduced a famous scattering term into the Liouville equation in an effort to explain the behavior of gases. However, his aim was to do more than explain gases: he wanted to **derive** the then-empirical laws of thermodynamics from the complete and quantitative laws of classical mechanics. Boltzmann succeeded in the most part in achieving these goals through his scattering term. The Boltzmann transport equation has a fascinating legacy that gave birth to the fields of statistical mechanics and quantum theory by introducing the concepts of entropy and irreversibility, and an "arrow of time". We are now ready to learn how this came about, and use these ideas for semiconductor electronics and photonics.

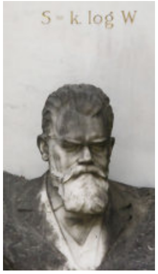


Fig. 21.5 Ludwig Boltzmann, the father of statistical mechanics and kinetic theory of gases. He formulated the formula for entropy S as a function of the degrees of freedom Ω: this formula $S = k_b \ln \Omega$, where k_b is the Boltzmann constant is inscribed on his gravestone. The concept of entropy permeates all branches of sciences, including communication systems. Boltzmann ended his life tragically by hanging himself in 1906.

From Fig. 21.6, to account for *scattering* events, we must also allow for $f(x,k,t)$ to *increase* by scattering "in" from nearby k' points on other trajectories, and for it to *decrease* by scattering "out":

$$f = f(x,k,t) = f(x - vdt, k - \frac{F}{\hbar}dt, t - dt) + (S_{in} - S_{out})dt, \tag{21.9}$$

where S_{in} and S_{out} are the in- and out-scattering *rates*, in units of 1/s. Taylor-expanding the first term on the right and rearranging, we get the 1D Boltzmann transport equation

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + \frac{F}{\hbar}\frac{\partial f}{\partial k} = S_{in} - S_{out}. \tag{21.10}$$

From Fig. 21.6 we see that the net in- and out-scattering rates depend on the occupation functions of the initial and final states. Let us denote $f(x,k,t) = f_k$ and $S(k \to k')$ as the individual scattering rate[13] from state $|k\rangle \to |k'\rangle$. The net rate of this process is proportional to f_k since there must be an electron in state k to scatter out of. The rate must also be proportional to $(1 - f_{k'})$ because if state k' was filled, another electron cannot scatter into it. Pauli's exclusion principle for electrons therefore requires the net scattering in and out rates to be

$$S_{in} = S(k' \to k) f_{k'}(1 - f_k), \text{ and } S_{out} = S(k \to k') f_k (1 - f_{k'}). \tag{21.11}$$

Now to track the occupation function of state $|k\rangle$, we must allow for in- and out-scattering to *all* other states $|k'\rangle$. Then, using Equations 21.11 in 21.10, the Boltzmann transport equation becomes:

$$\begin{aligned} &\frac{\partial f_k}{\partial t} + v\frac{\partial f_k}{\partial x} + \frac{F}{\hbar}\frac{\partial f_k}{\partial k} \\ &= \underbrace{\sum_{k'} [S(k' \to k) f_{k'}(1 - f_k) - S(k \to k') f_k (1 - f_{k'})]}_{\hat{C}f_k}. \end{aligned} \tag{21.12}$$

The left hand side called the **flow term** is identical to the Liouville equation. The right-hand side is called the **collision integral** and formally denoted as $\hat{C}f_k$. It accounts for all types of scattering and

---

[13] The individual scattering rate is found using Fermi's golden rule: $S(k \to k') = \frac{2\pi}{\hbar}|W_{k,k'}|^2 \delta(E_k - E_{k'} \pm \hbar\omega)$, where $W_{k,k'} = \langle k'|W(r)|k\rangle$ is the matrix element of the scattering potential $W(r)$ between states $|k\rangle$ and $|k'\rangle$ as was discussed in Chapter 20.

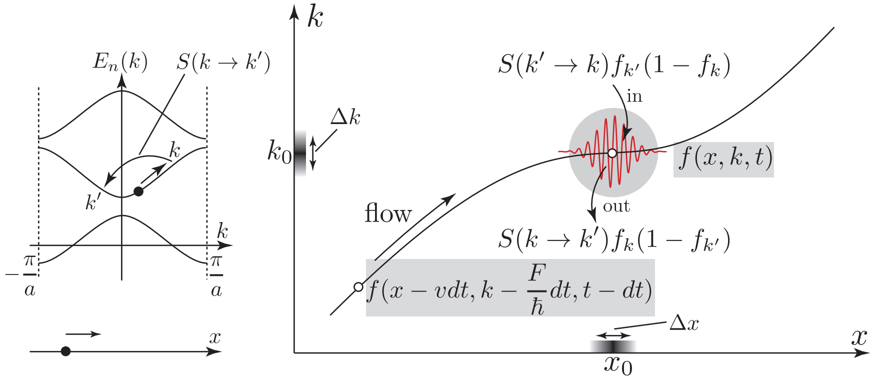


Fig. 21.6 Connecting the picture of scattering within the band of a semiconductor, with the picture of scattering in the phase–space. The Boltzmann collision integral accounts for in- and out-scattering from the Liouvillean ballistic trajectories of the wavepackets. At equilibrium, $f(x, k, t)$ is the Fermi–Dirac distribution function. Its non-equilibrium variation in the presence of scattering and external fields is obtained by solving the BTE. A scattering event from $k \to k'$ occurs at the same point in real space.

collisions, and the summation over all k' is converted to an integral $\sum_{k'}(...) \to \int \frac{dk'}{(2\pi/L)}(...)$. Solving the BTE in Equation 21.12 in principle gives us $f_k = f(x, k, t)$, the occupation probability of the wavepacket centered at $|k\rangle$ and x at time t in response to external electric fields or potential gradients, in the presence of all sorts of scattering events in the crystal[14]. But the BTE in its raw form is not that straightforward to solve in the most general case, because f_k also appears in a tangled form inside the sum (or collision integral) on the right. This is why the BTE is called an **integro-differential equation**. Before we start seeking solutions which will reveal a rich spectrum of phenomena, let us first generalize the BTE to higher dimensions by writing it as

$$\boxed{\frac{\partial f_{\mathbf{k}}}{\partial t} + \mathbf{v}_{\mathbf{k}} \cdot \nabla_{\mathbf{r}} f_{\mathbf{k}} + \frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f_{\mathbf{k}} = \hat{C} f_{\mathbf{k}},} \tag{21.13}$$

where on the left side, $\mathbf{v}_{\mathbf{k}} = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k})$ is the group velocity vector of state $\mathbf{k}$, and $\mathbf{F}$ is the external force acting on the electrons[15]. The collision integral on the right can be visualized using Fig. 21.7 (a), and classified into the three types of scattering events in that figure.

Fig. 21.7 (b) shows elastic scattering, when $|\mathbf{k}| = |\mathbf{k}'|$, and $E_{\mathbf{k}} = E_{\mathbf{k}'}$. An example of elastic scattering is when an electron in state $\mathbf{k}$ scatters from a charged dopant ion in the semiconductor into a state $\mathbf{k}'$. Now energy and momentum are *always* conserved in every scattering event *for the whole system*. But since the mass of the ion M is much larger than the electron m_e, $m_e \Delta v_e \sim M \Delta v_{ion} \implies \Delta v_{ion} \sim \frac{m_e}{M}\Delta v_e \approx 0$. This implies that to a very good approximation, the heavy ion can change the momentum of the electron without changing the electron's kinetic energy, making it an elastic process *for the electron*. The collision term for elastic scattering that tracks the occupation of state $\mathbf{k}$ is then the

[14] Knowledge of f_k gives us near magical powers, because we can then easily calculate the macroscopic charge (or heat, spin, ...) transport currents due to the electron via sums of the type $\sum_k \phi_k f_k$, where ϕ_k is the physical quantity of interest, as a result of all *microscopic* scattering events! f_k connects the microscopic processes to the macroscopic observables, and is found by solving the BTE.

[15] The internal forces due to the periodic atomic potential are already considered in the energy bandstructure $E(\mathbf{k})$, which is buried in $f_{\mathbf{k}}$. The term $\mathbf{F}$ is therefore only due to external forces such as the Lorentz force due to electric and magnetic fields.

difference of the incoming and outgoing fluxes

$$\hat{C}f_{\mathbf{k}} = \sum_{\mathbf{k}'}[S(\mathbf{k}'\to\mathbf{k})f_{\mathbf{k}'}(1-f_{\mathbf{k}}) - S(\mathbf{k}\to\mathbf{k}')f_{\mathbf{k}}(1-f_{\mathbf{k}'})]$$
$$\underbrace{S(\mathbf{k}'\to\mathbf{k}) = S(\mathbf{k}\to\mathbf{k}')}_{\text{elastic scattering}} \implies \hat{C}f_{\mathbf{k}} = \sum_{\mathbf{k}'} S(\mathbf{k}'\to\mathbf{k})(f_{\mathbf{k}'} - f_{\mathbf{k}}). \quad (21.14)$$

The equality $S(\mathbf{k}'\to\mathbf{k}) = S(\mathbf{k}'\to\mathbf{k}) = \frac{2\pi}{\hbar}|W_{\mathbf{k},\mathbf{k}'}|^2\delta(E_{\mathbf{k}} - E_{\mathbf{k}'})$ is guaranteed by Fermi's golden rule for elastic processes. This is because the scattering matrix element $\langle\mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle = (\langle\mathbf{k}|W(\mathbf{r})|\mathbf{k}'\rangle)^\star$, implying $|W_{\mathbf{k},\mathbf{k}'}|^2 = |W_{\mathbf{k}',\mathbf{k}}|^2$, and for the Dirac delta function, $\delta(+x) = \delta(-x)$ is an identity. Note that the sum (or integral) is over all $\mathbf{k}'$.

The collision term for inelastic scattering (Fig. 21.7 (c)) is

$$\hat{C}f_{\mathbf{k}} = \sum_{\mathbf{k}'}[S(\mathbf{k}'\to\mathbf{k})f_{\mathbf{k}'}(1-f_{\mathbf{k}}) - S(\mathbf{k}\to\mathbf{k}')f_{\mathbf{k}}(1-f_{\mathbf{k}'})]$$
$$S(\mathbf{k}\to\mathbf{k}') = \sum_{\mathbf{q}}[\underbrace{S(\mathbf{k},n_{\mathbf{q}}\to\mathbf{k}',n_{\mathbf{q}}+1)}_{\text{emission}\,\propto\, n_{\mathbf{q}}+1} + \underbrace{S(\mathbf{k},n_{\mathbf{q}}\to\mathbf{k}',n_{\mathbf{q}}-1)}_{\text{absorption}\,\propto\, n_{\mathbf{q}}}]$$
$$S(\mathbf{k}'\to\mathbf{k}) = \sum_{\mathbf{q}}[\underbrace{S(\mathbf{k}',n_{\mathbf{q}}\to\mathbf{k},n_{\mathbf{q}}+1)}_{\text{emission}\,\propto\, n_{\mathbf{q}}+1} + \underbrace{S(\mathbf{k}',n_{\mathbf{q}}\to\mathbf{k},n_{\mathbf{q}}-1)}_{\text{absorption}\,\propto\, n_{\mathbf{q}}}].$$
$$(21.15)$$

Here, $S(\mathbf{k},n_{\mathbf{q}}\to\mathbf{k}',n_{\mathbf{q}}+1)$ is the scattering rate of an electron from state $\mathbf{k}$ to state $\mathbf{k}'$ by the emission of a phonon. The energy $E_{\mathbf{k}'} = E_{\mathbf{k}} \pm \hbar\omega_{\mathbf{q}}$ is conserved for such emission and absorption processes for the combined system, and so is the momentum $\hbar\mathbf{k}' = \hbar\mathbf{k} \pm \hbar\mathbf{q}$. Because of the emission, the number of phonons $n_{\mathbf{q}}$ in mode $\mathbf{q}$ before scattering increases by one quantum, to $n_{\mathbf{q}}+1$ after the scattering event. In Chapter 16, Equation 16.18 we proved that the rates for emission and absorption are not the same, but related by the Maxwell–Boltzmann factor $S(\mathbf{k}\to\mathbf{k}') = S(\mathbf{k}'\to\mathbf{k}) \times \exp[(E_{\mathbf{k}} - E_{\mathbf{k}'})/k_bT]$. Note that though in Equation 21.15 we have written out both emission and absorption processes for in- and out-scattering from state $\mathbf{k}\to\mathbf{k}'$, only one of them is allowed for a given choice of $E_{\mathbf{k}}, E_{\mathbf{k}'}$ since either $E_{\mathbf{k}} > E_{\mathbf{k}'}$ or $E_{\mathbf{k}} < E_{\mathbf{k}'}$ for inelastic scattering, and only one pair survives. Using these properties, the collision integral for inelastic scattering may be written as

$$\hat{C}f_{\mathbf{k}} = \sum_{\mathbf{k}'} S(\mathbf{k}\to\mathbf{k}')[e^{\frac{E_{\mathbf{k}}-E_{\mathbf{k}'}}{k_bT}} f_{\mathbf{k}'}(1-f_{\mathbf{k}}) - f_{\mathbf{k}}(1-f_{\mathbf{k}'})], \quad (21.16)$$

which we recognize as correct also for elastic scattering, which is the special case when $E_{\mathbf{k}} = E_{\mathbf{k}'}$. For inelastic scattering, implicit in $S(\mathbf{k}\to\mathbf{k}')$ is the sum $\sum_{\mathbf{q}}(...)$ over phonon modes as seen in Equation 21.15.

For electron-electron scattering, we have from Fig. 21.7,

$$\hat{C}f_{\mathbf{k}} = \sum_{\mathbf{k}_1,\mathbf{k}',\mathbf{k}'_1} [S(\mathbf{k}',\mathbf{k}'_1\to\mathbf{k},\mathbf{k}_1)f_{\mathbf{k}'}f_{\mathbf{k}'_1}(1-f_{\mathbf{k}})(1-f_{\mathbf{k}_1})$$
$$-S(\mathbf{k},\mathbf{k}_1\to\mathbf{k}',\mathbf{k}'_1)f_{\mathbf{k}}f_{\mathbf{k}_1}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}'_1})], \quad (21.17)$$

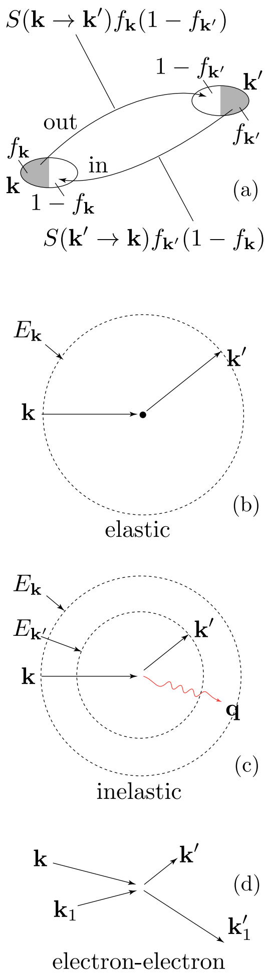


Fig. 21.7 Illustration of the scattering term of the Boltzmann equation (a), and the three types of scattering events: (b) electron-defect, (c) electron-phonon, and (d) electron-electron scattering.

in which to track the occupation probability $f_{\mathbf{k}}$ of state $\mathbf{k}$, we must sum the rates over all other initial states $\mathbf{k_1}$, and all final states $\mathbf{k}'$ and $\mathbf{k}'_1$. Thus, the sum converted to an integral will take the form $\sum_{\mathbf{k_1},\mathbf{k}',\mathbf{k}'_1}(...) \rightarrow \int \frac{d\mathbf{k_1}d\mathbf{k}'d\mathbf{k}'_1}{(2\pi)^{3d}}(...)$. The electron-electron scattering rate $S(\mathbf{k}',\mathbf{k}'_1 \rightarrow \mathbf{k},\mathbf{k_1})$ has the repulsive Coulomb interaction energy as the scattering potential, and respects the exchange process of fermions as was discussed in Chapter 20. With a better understanding of the microscopic nature of the collision term[16] of the BTE, let us now see one of the rather remarkable theorems first proven by Boltzmann.

21.4 H-theorem and entropy

In 1872, Boltzmann "derived", from Newton's laws of motion, the **second law of thermodynamics** (see Chapter 4), which states that spontaneous processes always **increase the entropy** of a system. He started with the question[17]: **how do collisions change the entropy of the electron system with time**? We will see now that collisions take a system from an excited state of low entropy to its equilibrium state, where the entropy is maximized. This will occur in a manner that the rate of change of entropy is positive (Boltzmann's H-theorem).

In fact the BTE will enable us to answer a far more general question: how do collisions and scattering determine the time evolution of *any* physical property of interest? Let us denote the physical property by $\phi_{\mathbf{k}}$, which can be a time-dependent function of the wavevector $\mathbf{k}$ *and* the occupation probability $f_{\mathbf{k}}$ of the BTE. For now, let us focus on the effect of collisions when there are no concentration gradients or forces, implying $(\mathbf{v_k}\cdot\nabla_{\mathbf{r}}f_{\mathbf{k}} + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}}f_{\mathbf{k}}) = 0$ in Equation 21.13. Denoting $\hat{C}f = \frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}}$, the time evolution of the macroscopic (measurable) value of the physical property $\rho_\phi(\mathbf{r},t) = \sum_{\mathbf{k}}\phi_{\mathbf{k}}f_{\mathbf{k}}(\mathbf{r},t)$ is given by

$$\frac{\partial\rho_\phi}{\partial t} = \frac{\partial}{\partial t}\sum_{\mathbf{k}}\phi_{\mathbf{k}}f_{\mathbf{k}}(\mathbf{r},t) = \sum_{\mathbf{k}}\frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}}\cdot(\phi_{\mathbf{k}} + f_{\mathbf{k}}\frac{\partial\phi_{\mathbf{k}}}{\partial f_{\mathbf{k}}}) = \sum_{\mathbf{k}}\frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}}\cdot\bar{\phi}_{\mathbf{k}}, \tag{21.18}$$

where[18] $\bar{\phi}_{\mathbf{k}} = \left(\phi_{\mathbf{k}} + f_{\mathbf{k}}\frac{\partial\phi_{\mathbf{k}}}{\partial f_{\mathbf{k}}}\right)$. So to find how collisions change the physical property $\phi_{\mathbf{k}}$, we must evaluate the sum (or integral) of Equation 21.18 for the appropriate property. For the collision term $(\partial f_{\mathbf{k}}/\partial t)|_{\text{coll.}}$ of Equation 21.18, consider the situation when defect and phonon scattering are absent. Then, electron-electron scattering of Equation 21.17 survives as the only term in the collision integral of the BTE. Neglecting electron spin, every electron-electron scattering event is elastic[19], and follows $S(\mathbf{k},\mathbf{k_1} \rightarrow \mathbf{k}',\mathbf{k}'_1) = S(\mathbf{k}',\mathbf{k}'_1 \rightarrow \mathbf{k},\mathbf{k_1})$. The rate of change of ρ_ϕ in Equation 21.18 then becomes

$$\frac{\partial\rho_\phi}{\partial t} = \sum_{\mathbf{k},\mathbf{k_1},\mathbf{k}',\mathbf{k}'_1} S(\mathbf{k}',\mathbf{k}'_1 \rightarrow \mathbf{k},\mathbf{k_1}) \times \bar{\phi}_{\mathbf{k}} \times$$

$$[f_{\mathbf{k}'}f_{\mathbf{k}'_1}(1-f_{\mathbf{k}})(1-f_{\mathbf{k}_1}) - f_{\mathbf{k}}f_{\mathbf{k}_1}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}'_1})]. \tag{21.19}$$

---

[16] The Boltzmann equation captures kinetic phenomena in the "dilute" limit when $nd^3 << 1$, when the average distance between electrons $1/n^{1/3}$ is far larger than the effective size of the wavepacket d. The other end of kinetic equations occurs in the dense limit $nd^3 >> 1$, which is called the **Vlasov equation**, used in the study of plasmas. The collisions in BTE follow the rules of quantum mechanics, and conserve energy and momentum on the whole and are in this sense "deterministic". The BTE-like variants: **Langevin** and **Fokker–Planck** equations introduce fictitious mathematical randomizing processes without regard to the physical scattering mechanisms. These models are used to study noise and fluctuations in semiconductors, but we do not discuss them further.

[17] Note that Boltzmann's original analysis was not for electrons, but for molecules of a gas, following the laws of *classical* mechanics. Neither the electron, nor their quantum mechanics was known at the time! Boltzmann's arguments have been adapted in the discussion of this section for electrons and their complete quantum statistics.

[18] Note that this is an example of taking the *moments* of the BTE. This specific equilibrium case is generalized to non-equilibrium situations in Section 21.6. Also note that if $\phi_{\mathbf{k}}$ depends on $\mathbf{k}$ and on $f_{\mathbf{k}}$ (and hence on time), then $\bar{\phi}_{\mathbf{k}} = (\phi_{\mathbf{k}} + f_{\mathbf{k}}\frac{\partial\phi_{\mathbf{k}}}{\partial f_{\mathbf{k}}})$. If $\phi_{\mathbf{k}}$ depends on $\mathbf{k}$ and not on $f_{\mathbf{k}}$, then $\bar{\phi}_{\mathbf{k}} = \phi_{\mathbf{k}}$.

[19] Electron-electron scattering is analogous to the inter-molecular collisions in a gas. Spin is an internal degree of freedom of the electron, analogous to the internal degrees of freedom of molecules, such as rotational or vibrational modes.

The summation is over all four $\mathbf{k}$-wavevectors indicated in Fig. 21.7 (d). By solving Exercise 21.2 with a mathematical trick interchanging the four $\mathbf{k}'s$ using physical arguments, you can show (as Boltzmann did), that the time evolution of the property ρ_ϕ of the electron system is then given by the somewhat non-intuitive relation that holds for **any** $\phi_{\mathbf{k}}$:

$$\begin{aligned}\frac{\partial \rho_\phi}{\partial t} = -\frac{1}{4}\sum_{\mathbf{k},\mathbf{k_1},\mathbf{k}',\mathbf{k}'_1} & S(\mathbf{k},\mathbf{k_1}\to\mathbf{k}',\mathbf{k}'_1)\times \\ & (\bar{\phi}_{\mathbf{k}}+\bar{\phi}_{\mathbf{k_1}}-\bar{\phi}_{\mathbf{k}'}-\bar{\phi}_{\mathbf{k}'_1})\times \\ [f_{\mathbf{k}}f_{\mathbf{k_1}}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}'_1}) & - f_{\mathbf{k}'}f_{\mathbf{k}'_1}(1-f_{\mathbf{k}})(1-f_{\mathbf{k_1}})].\end{aligned} \tag{21.20}$$

For a demonstration of the deep physics hiding in Equation 21.20, let us first choose that particular $\phi_{\mathbf{k}}$ such that the corresponding ρ_ϕ represents the **entropy** of the electron system. Recall from Chapter 4, Equation 4.3 that the entropy of an electron state $\mathbf{k}$ is given in terms of its occupation function $f_{\mathbf{k}}$ by $\mathcal{S}_{\mathbf{k}} = -k_b[f_{\mathbf{k}}\ln f_{\mathbf{k}} + (1-f_{\mathbf{k}})\ln(1-f_{\mathbf{k}})]$ due to its fermionic statistics, in a general non-equilibrium situation[20]. To find the rate of change of the entropy of the entire distribution $\rho_\phi(\mathbf{r},t) = -k_b\cdot\sum_{\mathbf{k}}[f_{\mathbf{k}}\ln f_{\mathbf{k}} + (1-f_{\mathbf{k}})\ln(1-f_{\mathbf{k}})]$, direct substitution verifies that we should choose the function

$$\phi_{\mathbf{k}} = \frac{\mathcal{S}_{\mathbf{k}}}{f_{\mathbf{k}}} = -k_b\sum_{\mathbf{k}}\big(\ln f_{\mathbf{k}} + \frac{1-f_{\mathbf{k}}}{f_{\mathbf{k}}}\ln(1-f_{\mathbf{k}})\big). \tag{21.21}$$

Then, from Equation 21.18 we have

$$\bar{\phi}_{\mathbf{k}} = (\phi_{\mathbf{k}} + f_{\mathbf{k}}\frac{\partial \phi_{\mathbf{k}}}{\partial f_{\mathbf{k}}}) = k_b\cdot\ln(\frac{1-f_{\mathbf{k}}}{f_{\mathbf{k}}}), \tag{21.22}$$

and using it in Boltzmann's form in Equation 21.20, we obtain

$$\begin{aligned}\frac{\partial \rho_\phi}{\partial t} = -\frac{k_b}{4}\sum_{\mathbf{k},\mathbf{k_1},\mathbf{k}',\mathbf{k}'_1} & S(\mathbf{k},\mathbf{k_1}\to\mathbf{k}',\mathbf{k}'_1)\times \\ [f_{\mathbf{k}}f_{\mathbf{k_1}}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}'_1}) & - f_{\mathbf{k}'}f_{\mathbf{k}'_1}(1-f_{\mathbf{k}})(1-f_{\mathbf{k_1}})]\times \\ [\ln(\frac{1-f_{\mathbf{k}}}{f_{\mathbf{k}}}\cdot\frac{1-f_{\mathbf{k_1}}}{f_{\mathbf{k_1}}}) & - \ln(\frac{1-f_{\mathbf{k}'}}{f_{\mathbf{k}'}}\cdot\frac{1-f_{\mathbf{k}'_1}}{f_{\mathbf{k}'_1}})] \\ \implies \frac{\partial \rho_\phi}{\partial t} = +\frac{k_b}{4}\sum_{\mathbf{k},\mathbf{k_1},\mathbf{k}',\mathbf{k}'_1} & S(\mathbf{k},\mathbf{k_1}\to\mathbf{k}',\mathbf{k}'_1)\times \\ [\underbrace{f_{\mathbf{k}}f_{\mathbf{k_1}}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}'_1})}_{x_2} & - \underbrace{f_{\mathbf{k}'}f_{\mathbf{k}'_1}(1-f_{\mathbf{k}})(1-f_{\mathbf{k_1}})}_{x_1}]\times \\ [\underbrace{\ln[f_{\mathbf{k}}f_{\mathbf{k_1}}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}'_1})]}_{\ln(x_2)} & - \underbrace{\ln[f_{\mathbf{k}'}f_{\mathbf{k}'_1}(1-f_{\mathbf{k}})(1-f_{\mathbf{k_1}})]}_{\ln(x_1)}] \\ \implies \boxed{\frac{\partial \rho_\phi}{\partial t} = +\frac{k_b}{4}\sum_{\mathbf{k}'s} S(...)\times(x_2-x_1)\times\ln(\frac{x_2}{x_1}).} &\end{aligned} \tag{21.23}$$

[20] The script $\mathcal{S}$ is used to distinguish entropy from the scattering rate S.

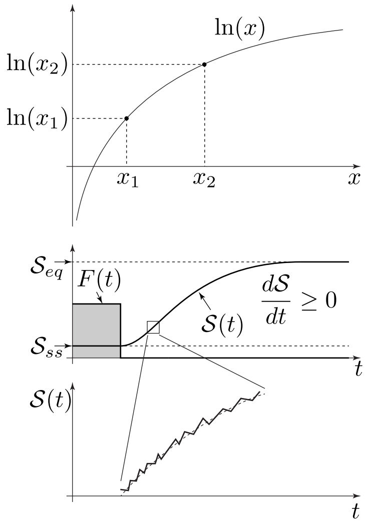


Fig. 21.8 Visual proof that $(x_2 - x_1) \times \ln(\frac{x_2}{x_1}) \geq 0$, and as a result, the entropy $\rho_\phi = \mathcal{S}(t)$ increases with time. This is analogous to the second law of thermodynamics, applied to the electron system.

Since k_b and the scattering rate $S(...)$ are positive, whether the macroscopic entropy of the electron system increases or decreases with time hinges on the sign of the factor $(x_2 - x_1) \times \ln(\frac{x_2}{x_1})$ in Equation 21.23. Because the natural log is a monotonic function, for any choice of the pair (x_2, x_1) we have the identity $(x_2 - x_1) \times \ln(\frac{x_2}{x_1}) \geq 0$, as seen in Fig. 21.8. This is why the **entropy of the electron gas left to itself with no external fields will either increase, or stay constant with time due to collisions**. This is **Boltzmann's H-theorem** for the rate of change of entropy due to collisions[21]. A typical situation is shown in Fig. 21.8: when an external force (say due to an electric field) is turned off, the entropy of the electron system spontaneously increases from the steady state value $\mathcal{S}_{ss}$ to its equilibrium value $\mathcal{S}_{eq}$ in a manner that the slope is positive. In the next section, we use the fact that equilibrium is defined when the rate of entropy production ceases: this happens when $x_2 = x_1$ in the Equation 21.23. This fact, with the fundamental conservation laws of physics will allow us to derive the equilibrium $f_{\mathbf{k}}$, which will turn out to be the Fermi–Dirac distribution $f_0 = 1/(\exp[(E - \mu)/k_b T] + 1)$ for electrons. Fig. 21.8 shows a zoomed in plot of the entropy function showing that there may be regions where the local $d\mathcal{S}(t)/dt < 0$. This is because correlations can violate the approximation of independent particles of the Liouville equation that led to the BTE. The entropy on the coarser scale must increase, maximize, and stop changing at equilibrium.

The startling conclusion about the unidirectional, irreversible path a physical system spontaneously takes towards equilibrium, which is concluded here for electrons (that follow quantum statistics and quantum mechanics) remains unchanged for the molecules of a gas of classical particles subject to Newton's laws of motion and classical collisions. In fact, this is Boltzmann's original "derivation" of the second law of thermodynamics, which explains why, for example, a gas that spontaneously expands from a small to a large volume does not return back spontaneously to the small volume. In a similar way, electrons or holes that spontaneously diffuse out from a region of a semiconductor will not spontaneously return to that volume: this process is irreversible. Associated with every spontaneous process therefore is **an arrow of time that always points towards increasing entropy: dictating that spontaneous physical processes are irreversible**. A "video" of the process can unambiguously identify if it is allowed by the physical laws of nature and distinguish the natural from the unnatural.

What may seem rather mysterious is: where exactly did the irreversibility creep in? The microscopic laws governing the dynamics of systems are Schrödinger's equation in quantum mechanics, and Newton's laws in classical mechanics. Both are completely reversible because replacing $t \to -t$ does not change the equations and solutions. Scattering processes are also completely reversible for the whole system. Boltzmann's initial explanation was that after each scattering event, the particles are free to choose any one of the many final states that satisfy the conservation laws (of momentum, energy, angular mo-

[21] Boltzmann called $H = f \ln(f)$, effectively the negative of what we call entropy today and showed that H *decreases* with time. This is why it is referred to as the H-theorem. There are no $1 - f$ factors; Boltzmann's f was the *classical* occupation function of molecules of a gas in the $(\mathbf{r}, \mathbf{p})$ phase space. Using the conservation arguments discussed in the next section, he derived the Maxwell–Boltzmann distribution of gas molecules at equilibrium. Historically several attempts have been made to find violations of the Boltzmann H-theorem, famously resulting in the formulation of quantum mechanics by Max Planck, and several advances in mathematics resulting in the proof of the Poincare conjecture. See Exercise 21.6 for further exploration.

mentum, etc.). He coined the word molecular chaos (*stosszahlansatz*) for this phenomena. Though this *interpretation* has been hotly contested and argued till this day, Boltzmann's overall conclusion that equilibrium is the state of highest entropy, and any system left to itself will irreversibly end up in the equilibrium state via collisions and scattering is not just unscathed, it forms the foundation of thermodynamics, statistical mechanics and quantum mechanics[22]. See Exercise 21.3 for further thoughts on this matter. Let us turn our attention to the summit of entropy: the equilibrium state.

21.5 Equilibrium distribution

The solution of the BTE gives the occupation probability $f_{\mathbf{k}}$ of each $\mathbf{k}$ state for particles that have any general bandstructure of the form $E(\mathbf{k})$, in *both* equilibrium, and non-equilibrium conditions. In Chapter 4 we found by maximizing the entropy and enforcing quantum statistics that the equilibrium distributions of fermions and bosons are given by the Fermi–Dirac and Bose–Einstein functions respectively. It should be possible to derive these functions directly from the BTE too. We show now[23] that it indeed is the case, and is a consequence of the fundamental conservation laws.

Let us consider the equilibrium case when Equation 21.20 holds. Unlike the last section where we chose the specific $\phi_{\mathbf{k}}$ for entropy density, we now choose a $\phi_{\mathbf{k}}$ that is *not* a function of $f_{\mathbf{k}}$, but $\mathbf{k}$ alone, implying $\bar{\phi}_{\mathbf{k}} = \left(\phi_{\mathbf{k}} + f_{\mathbf{k}} \frac{\partial \phi_{\mathbf{k}}}{\partial f_{\mathbf{k}}}\right) = \phi_{\mathbf{k}}$ from Equation 21.22, and therefore

$$\frac{\partial \rho_\phi}{\partial t} = -\frac{1}{4} \sum_{\mathbf{k},\mathbf{k}_1,\mathbf{k}',\mathbf{k}_1'} S(\mathbf{k},\mathbf{k}_1 \to \mathbf{k}',\mathbf{k}_1') \times (\phi_{\mathbf{k}} + \phi_{\mathbf{k}_1} - \phi_{\mathbf{k}'} - \phi_{\mathbf{k}_1'}) \times \\ [f_{\mathbf{k}} f_{\mathbf{k}_1}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}_1'}) - f_{\mathbf{k}'} f_{\mathbf{k}_1'}(1-f_{\mathbf{k}})(1-f_{\mathbf{k}_1})]. \tag{21.24}$$

The LHS of this equation is zero for a macroscopic physical quantity ρ_ϕ that is conserved in collisions, because then it does not change with time. What are such quantities? The number of electrons is conserved in a scattering event: for example in electron-electron scattering, there are two electrons before and two after the scattering event. This is captured by $\phi_{\mathbf{k}} = \phi_{\mathbf{k}_1} = \phi_{\mathbf{k}'} = \phi_{\mathbf{k}_1'} = 1$, and $\phi_{\mathbf{k}} + \phi_{\mathbf{k}_1} = \phi_{\mathbf{k}'} + \phi_{\mathbf{k}_1'}$. Similarly, choosing $\phi_{\mathbf{k}} = E_{\mathbf{k}}$, $\phi_{\mathbf{k}_1} = E_{\mathbf{k}_1}$, $\phi_{\mathbf{k}'} = E_{\mathbf{k}'}$, and $\phi_{\mathbf{k}_1'} = E_{\mathbf{k}_1'}$ again requires $\phi_{\mathbf{k}} + \phi_{\mathbf{k}_1} = \phi_{\mathbf{k}'} + \phi_{\mathbf{k}_1'}$ because $E_{\mathbf{k}} + E_{\mathbf{k}_1} = E_{\mathbf{k}'} + E_{\mathbf{k}_1'}$ is the statement of energy conservation in the collision. Corresponding statements hold for all other physical properties (such as momentum and angular momentum) that are conserved in the collision process. Note that the above quantities are conserved in collisions even when the system is *not* in equilibrium. Therefore physical quantities $\phi_{\mathbf{k}}$ that are conserved in collisions and make $\frac{\partial \rho_\phi}{\partial t} = 0$ exhibit the additive feature that $\sum_{\text{before}} \phi_{\mathbf{k}} = \sum_{\text{after}} \phi_{\mathbf{k}'}$.

---

[22] Richard Tolman in his book **The Principles of Statistical Mechanics** has this to say about Boltzmann's H-theorem: *"The derivation of this theorem and the appreciation of its significance may be regarded as one of the greatest achievements of physical science"*.

[23] We tackle the case of fermions here, and you do the same for bosons in Exercise 21.8.

In the last section we found that equilibrium is characterized by zero entropy production $\frac{\partial \mathcal{S}}{\partial t} = 0$, which strictly requires

$$\begin{aligned} f_{\mathbf{k}} f_{\mathbf{k}_1}(1-f_{\mathbf{k}'})(1-f_{\mathbf{k}'_1}) &= f_{\mathbf{k}'} f_{\mathbf{k}'_1}(1-f_{\mathbf{k}})(1-f_{\mathbf{k}_1}) \\ \implies \frac{f_{\mathbf{k}}}{1-f_{\mathbf{k}}} \cdot \frac{f_{\mathbf{k}_1}}{1-f_{\mathbf{k}_1}} &= \frac{f_{\mathbf{k}'}}{1-f_{\mathbf{k}'}} \cdot \frac{f_{\mathbf{k}'_1}}{1-f_{\mathbf{k}'_1}} \\ \implies \ln(\frac{f_{\mathbf{k}}}{1-f_{\mathbf{k}}}) + \ln(\frac{f_{\mathbf{k}_1}}{1-f_{\mathbf{k}_1}}) &= \ln(\frac{f_{\mathbf{k}'}}{1-f_{\mathbf{k}'}}) + \ln(\frac{f_{\mathbf{k}'_1}}{1-f_{\mathbf{k}'_1}}). \end{aligned} \tag{21.25}$$

Evidently, the dimensionless functional form $\ln(\frac{f_{\mathbf{k}}}{1-f_{\mathbf{k}}})$ has the additive form for quantities that are conserved in collisions $\sum_{\text{before}} \phi_{\mathbf{k}} = \sum_{\text{after}} \phi_{\mathbf{k}'}$. If we set $\ln(\frac{f_{\mathbf{k}}}{1-f_{\mathbf{k}}}) = c \cdot E_{\mathbf{k}}$ where c is a constant of inverse energy dimensions, then we obtain $f_{\mathbf{k}} = 1/[\exp[c \cdot E_{\mathbf{k}}] + 1]$. Though this form looks similar to the Fermi–Dirac distribution, it cannot be quite correct because energy is not the *only* conserved quantity in the collision. In electron-electron collisions, the quantities that are conserved are particle number ($\phi_{\mathbf{k}} = 1$), momentum ($\phi_{\mathbf{k}} = \mathbf{p}$), and energy ($\phi_{\mathbf{k}} = E_{\mathbf{k}}$). Then the quantity that must be *simultaneously* conserved is the dimensionless linear combination $a \cdot 1 + \mathbf{b} \cdot \mathbf{p} + c \cdot E$, where a, c are scalar constants, and $\mathbf{b}$ is a vector constant. The correct form of the equilibrium distribution function is then

$$\ln(\frac{f_{\mathbf{k}}}{1-f_{\mathbf{k}}}) = a + \mathbf{b} \cdot \mathbf{p} + c \cdot E \implies f_{\mathbf{k}} = \frac{1}{\exp\left[\frac{E_{\mathbf{k}} - \mathbf{u} \cdot \mathbf{p} - \mu}{k_b T}\right] + 1}. \tag{21.26}$$

From requirements of entropy and particle number, the constants are identified to be $a = \mu / k_b T$, where μ is the chemical potential as a measure of the particle number, $c = -1/k_b T$, and $\mathbf{b} = \mathbf{u}/k_b T$, where $\mathbf{u}$ is the drift velocity of the particle. Now in a semiconductor or a metal, an electron will collide every now and then with a stationary heavy donor atom or defect in an elastic manner such that the electron energy is conserved, but it's momentum is not. Thus we can remove the momentum conservation requirement if we want the distribution function of *electrons alone*, and recover $f^0_{\mathbf{k}} = 1/(\exp\left[\frac{E_{\mathbf{k}} - \mu}{k_b T}\right] + 1)$, the form of the Fermi–Dirac distribution that is correct for this situation[24].

We thus arrive at a **very important result: the solution of the BTE for electrons at thermal equilibrium is the Fermi–Dirac distribution function**, which we label as $f^0_{\mathbf{k}}$. This is a remarkable statement, because it is independent of the fact that the electrons may be colliding with various kinds of defects, scattering off of lattice vibrations of several kinds, or even thermally excited from one band to another. Though several scattering events are responsible to take electrons to their equilibrium distribution, **the equilibrium distribution itself is independent of the nature or type of the scattering events**. What the nature of scattering events *do* control are the dynamical and steady-state properties of the electron population $f_{\mathbf{k}}$ when it is taken away, or brought back towards equilibrium. For example, the scattering rates

[24] Boltzmann applied this technique to obtain the distribution of velocities of gas molecules of mass M as $f(v) = (\frac{M}{2\pi k_b T})^{\frac{3}{2}} 4\pi v^2 e^{-\frac{Mv^2}{2k_b T}}$ from Newton's laws. See Exercise 21.7.

determine exactly how long it takes for $f_{\mathbf{k}}$ to reach the equilibrium distribution $f^0_{\mathbf{k}}$ when an external perturbation is removed. Scattering processes also determine how long it takes for the electron system to go from the equilibrium $f^0_{\mathbf{k}}$ to a steady-state $f^{ss}_{\mathbf{k}}$ when a perturbation is turned on, and the precise steady state value $f^{ss}_{\mathbf{k}}$ it can reach. These deviations from equilibrium are discussed in the next section. In the rest of this section, we discuss a few more properties of the equilibrium distribution and microscopic processes that occur in equilibrium[25].

Principle of detailed balance: Equilibrium is defined by the condition when the RHS of the BTE is zero, i.e., $\hat{C}f_{\mathbf{k}} = 0$. For general inelastic scattering given by Equation 21.15, this requires the equilibrium distribution $f_{\mathbf{k}} = f^0_{\mathbf{k}}$ to satisfy

$$\begin{aligned}
&\sum_{\mathbf{k}'}[S(\mathbf{k}'\to\mathbf{k})f^0_{\mathbf{k}'}(1-f^0_{\mathbf{k}}) - S(\mathbf{k}\to\mathbf{k}')f^0_{\mathbf{k}}(1-f^0_{\mathbf{k}'})] = 0\\
&\implies S(\mathbf{k}\to\mathbf{k}')f^0_{\mathbf{k}}(1-f^0_{\mathbf{k}'}) = S(\mathbf{k}'\to\mathbf{k})f^0_{\mathbf{k}'}(1-f^0_{\mathbf{k}})\\
&\implies \boxed{\frac{S(\mathbf{k}\to\mathbf{k}')}{S(\mathbf{k}'\to\mathbf{k})} = \frac{1-f^0_{\mathbf{k}}}{f^0_{\mathbf{k}}}\cdot\frac{f^0_{\mathbf{k}'}}{1-f^0_{\mathbf{k}'}} = e^{\frac{E_{\mathbf{k}}-E_{\mathbf{k}'}}{k_bT}}.}
\end{aligned} \tag{21.27}$$

In the above equation every single term inside the sum must be identically zero. The rate of *every* scattering event is *exactly* counterbalanced by the reverse process. This is the **principle of detailed balance**[26]. If it were not so, and say $S(\mathbf{k}\to\mathbf{k}')f^0_{\mathbf{k}}(1-f^0_{\mathbf{k}'}) > S(\mathbf{k}'\to\mathbf{k})f^0_{\mathbf{k}'}(1-f^0_{\mathbf{k}})$, the occupation of state $f^0_{\mathbf{k}}$ will decrease with time as particles scatter out faster than they scatter into this state, and $f^0_{\mathbf{k}'}$ will correspondingly increase, making $f_{\mathbf{k}} \neq f^0_{\mathbf{k}}$ and violating equilibrium.

Elastic scattering rates: The scattering rate $S(\mathbf{k}'\to\mathbf{k})$ and its reverse $S(\mathbf{k}\to\mathbf{k}')$ given by Fermi's golden rule are *independent* of the occupation functions $f_{\mathbf{k}}$ and $f_{\mathbf{k}'}$ of electrons, but depend on the nature of scattering mechanisms. From the boxed Equation 21.27 it is seen that if $E_{\mathbf{k}} = E_{\mathbf{k}'}$ and the scattering is elastic, $S(\mathbf{k}\to\mathbf{k}') = S(\mathbf{k}'\to\mathbf{k})$. The collision term of BTE for elastic scattering then becomes[27]

$$\begin{aligned}
&\frac{\partial f_{\mathbf{k}}}{\partial t}\Big|^{\text{elastic}}_{\text{coll.}} = \sum_{\mathbf{k}'}[S(\mathbf{k}'\to\mathbf{k})f_{\mathbf{k}'}(1-f_{\mathbf{k}}) - S(\mathbf{k}\to\mathbf{k}')f_{\mathbf{k}}(1-f_{\mathbf{k}'})]\\
&\implies \boxed{\frac{\partial f_{\mathbf{k}}}{\partial t}\Big|^{\text{elastic}}_{\text{coll.}} = \sum_{\mathbf{k}'}S(\mathbf{k}'\to\mathbf{k})[f_{\mathbf{k}'} - f_{\mathbf{k}}].}
\end{aligned} \tag{21.28}$$

Scattering of electrons by charged ions and various static defects is an elastic process, as will be discussed in Chapter 23. This simple form of the collision term will allow a solution of the BTE to evaluate their effect of such scattering on electron transport via the definition of a relaxation time, discussed in the next section. Note in the above boxed equation the coefficient of $f_{\mathbf{k}}$ on the RHS is a relaxation time of the

[25] An extended preview of the BTE was provided in Chapter 16, Section 16.6. I recommend reviewing that section before proceeding in this chapter.

[26] Detailed balance results from statistical mechanics. It should not be confused with microscopic reversibility. The principle of microscopic reversibility is a purely quantum mechanical result with no connection to statistical mechanics. It states that when all particles are considered, the scattering rate in one direction is identical to its reverse, because of the time-reversibility of the Schrödinger equation. For example, for elastic scattering $S(\mathbf{k}\to\mathbf{k}') = S(\mathbf{k}'\to\mathbf{k})$ is microscopic reversibility, and *not* detailed balance.

[27] The boxed expression would have been obtained if the Pauli exclusion principle was not enforced in the scattering. Can you explain why it is so, since the Pauli exclusion principle can not be violated for electrons!

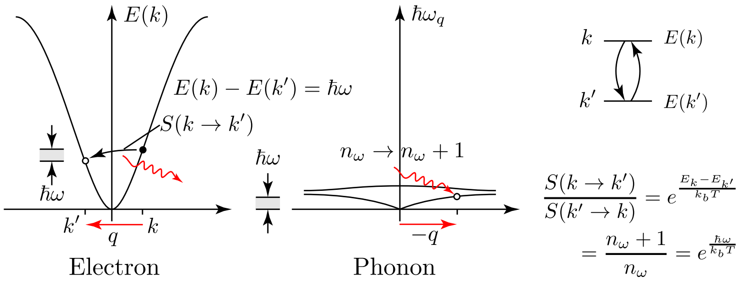

$E(k)$
$E(k) - E(k') = \hbar\omega$
$S(k \to k')$
$\hbar\omega$
k'
q
k
Electron
$\hbar\omega$
$\hbar\omega_q$
$n_\omega \to n_\omega + 1$
$-q$
Phonon
k
$E(k)$
k'
$E(k')$
$\frac{S(k \to k')}{S(k' \to k)} = e^{\frac{E_k - E_{k'}}{k_b T}}$
$= \frac{n_\omega + 1}{n_\omega} = e^{\frac{\hbar\omega}{k_b T}}$

Fig. 21.9 Detailed balance in electron-phonon scattering events.

form $\frac{1}{\tau(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k}' \to \mathbf{k})$, making the RHS linear in the unknown $f_{\mathbf{k}}$.

Inelastic scattering rates: Using the boxed Equation 21.27, the collision term of the BTE for inelastic scattering events is

$$\boxed{\frac{\partial f_{\mathbf{k}}}{\partial t}\Big|_{\text{coll.}}^{\text{inelastic}} = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')[e^{\frac{E_{\mathbf{k}'} - E_{\mathbf{k}}}{k_b T}} f_{\mathbf{k}'}(1 - f_{\mathbf{k}}) - f_{\mathbf{k}}(1 - f_{\mathbf{k}'})].} \tag{21.29}$$

The collision term does not simplify further for inelastic scattering. Because the RHS is in general a non-linear function of $f_{\mathbf{k}}$, a relaxation time cannot be strictly defined, except for special cases[28]. Fig. 21.9 depicts an inelastic scattering process for the electron which transitions from state $k \to k'$ of energies $E(k)$ and $E(k')$ which are different by $\hbar\omega$. The energy $E(k) - E(k') = \hbar\omega$ and momentum $\hbar k - \hbar k' = \hbar q$ lost by the electron must be gained by the object with which the electron collides, since the total energy and momentum of the (electron + scatterer) system is conserved. Consider the case when the electron scatters from phonons. The electron bandstructure $E(k)$ in Fig. 21.9 dictates the possible $\hbar\omega$ and $\hbar q$ for electron. The corresponding energy dispersion $\hbar\omega_q$ for phonons depends on the crystal structure, masses of the atoms, and the chemical bond strengths, when pictured as a periodic mass-spring system. The resulting phonon energy dispersion $\hbar\omega_q$ dictates which energies of phonons are available at particular wavevectors q. Only those electron scattering processes that match the phonon energy and momentum relation are allowed; this is treated in detail in Chapter 22.

For the scattering process shown, the number of phonons[29] in the mode n_ω increases by one, to $n_\omega + 1$. The scattering rate for phonon *emission* by electrons is proportional to $n_\omega + 1$, whereas the scattering rate for phonon *absorption* is proportional to n_ω, hence satisfying Equation 21.27. The factor of 1 in emission depicts the spontaneous part of emission, and the proportionality to n_ω the stimulated part of emission. Absorption of phonons is always stimulated. Note that the RHS of BTE in Equations 21.28 and 21.29 are manifestly non-equlibrium terms, but the arguments of the physics of equilibrium has baked into it several layers of physics, which will facilitate its solution[30].

[28] For example if $E_k - E_{k'} = \hbar\omega \gg k_b T$, the scattering term approximates to $\hat{C} f_k \approx \sum_{k'} S(k \to k') e^{\frac{\hbar\omega}{k_b T}} (f_{k'} - f_{k'} f_k)$, to which a relaxation time approximation may be crudely applied. This physical condition occurs for polar-optical phonon scattering in wide-bandgap semiconductors, where $\hbar\omega \gg k_b T$ holds at room temperature. Note that the elastic case is a special case of inelastic scattering when $E(k) = E(k')$.

[29] For phonons in equilibrium at a temperature T, the distribution function is $n_\omega^0 = 1/(\exp[\hbar\omega_q / k_b T] - 1)$, the Bose–Einstein distribution.

[30] Note that the argument of $\hbar\omega$ for a phonon – could be replaced by a photon, or other bosonic quanta with a corresponding dispersion. In other words, the concepts of absorption, spontaneous emission, and stimulated emission appear whenever fermionic (e.g. electron) systems interact with bosonic (e.g. photons, phonons) systems. Such interactions play a critical role in electron transport in crystals as well as in electron-photon or light-matter interactions.

Non-uniform equilibrium: We can coax out more from the BTE about the nature of the equilibrium distribution function. Now we show that the most general equilibrium distribution takes the form

$$\boxed{f^0_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\exp[\frac{E_c(\mathbf{r})+E(\mathbf{k})-E_F}{k_b T}]+1}, \text{ where } \mathbf{F} = -\nabla_{\mathbf{r}} E_c(\mathbf{r}).} \tag{21.30}$$

At equilibrium, we will strictly have $\nabla_{\mathbf{r}} E_F = 0$, *and* $\nabla_{\mathbf{r}} T = 0$. This means the Fermi level and the temperature must be constant in space, but the conduction band edge $E_c(\mathbf{r})$ can vary in space. At equilibrium $\mathbf{F} = -\nabla_{\mathbf{r}} E_c(\mathbf{r})$ is the *internal* force on an electron: it makes the electron roll down a gradient in the conduction band edge $E_c(\mathbf{r})$. An example of the non-uniform equilibrium condition encountered in semiconductor quantized structures is shown in Fig. 21.10.

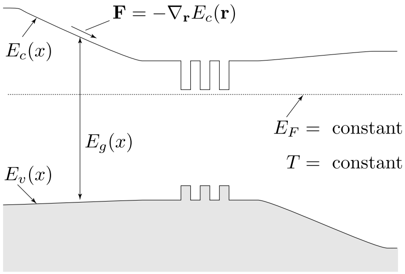


Fig. 21.10 Nonuniform equilibrium in a semiconductor graded-index separate confinement heterostructure (GRIN-SCH) quantum well laser active region.

To prove Equation 21.30, let us assume the more general form in which E_c, E_F, T are allowed to vary in space and time in the form:

$$f_{\mathbf{k}}(\mathbf{r},t) = \frac{1}{e^{\eta(\mathbf{r},\mathbf{k},t)}+1}, \text{ where } \eta(\mathbf{r},\mathbf{k},t) = \frac{E_c(\mathbf{r},t)+E(\mathbf{k})-E_F(\mathbf{r},t)}{k_b T(\mathbf{r},t)}. \tag{21.31}$$

Substituting this form into Equations 21.28 and 21.29 shows that it satisfies the requirement for equilibrium that the RHS of the BTE $\hat{C}\{f_{\mathbf{k}}(\mathbf{r},t)\} = 0$. But this is not enough: for $f_{\mathbf{k}}(\mathbf{r},t)$ to be an equilibrium distribution, it must also make the LHS of the BTE$= 0$:

$$\begin{aligned} &\underbrace{\frac{\partial f_{\mathbf{k}}(\mathbf{r},t)}{\partial t}}_{\frac{\partial f_{\mathbf{k}}(\mathbf{r},t)}{\partial \eta}\cdot\frac{\partial \eta}{\partial t}} + \mathbf{v}\cdot\underbrace{\nabla_{\mathbf{r}} f_{\mathbf{k}}(\mathbf{r},t)}_{\frac{\partial f_{\mathbf{k}}(\mathbf{r},t)}{\partial \eta}\cdot\nabla_{\mathbf{r}}\eta} + \frac{\mathbf{F}}{\hbar}\cdot\underbrace{\nabla_{\mathbf{k}} f_{\mathbf{k}}(\mathbf{r},t)}_{\frac{\partial f_{\mathbf{k}}(\mathbf{r},t)}{\partial \eta}\cdot\nabla_{\mathbf{k}}\eta} = 0 \\ &\implies \frac{\partial f_{\mathbf{k}}(\mathbf{r},t)}{\partial \eta}[\frac{\partial \eta}{\partial t} + \mathbf{v}\cdot\nabla_{\mathbf{r}}\eta + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}}\eta] = 0. \end{aligned} \tag{21.32}$$

Since $\partial f_{\mathbf{k}}(\mathbf{r},t)/\partial\eta \neq 0$, the term in the square bracket must be zero. Noting that $\nabla_{\mathbf{k}}\eta(\mathbf{r},\mathbf{k},t) = (\nabla_{\mathbf{k}} E(\mathbf{k}))/k_b T(\mathbf{r},t) = (\hbar\mathbf{v})/k_b T(\mathbf{r},t)$, and using the condition for equilibrium that all time dependences must vanish implies $\mathbf{v}\cdot[\mathbf{F} + k_b T(\mathbf{r})\nabla_{\mathbf{r}}\eta(\mathbf{r},\mathbf{k})] = 0$, which, when written as

$$\underbrace{[\mathbf{F} + \nabla_{\mathbf{r}} E_c(\mathbf{r})]}_{=0} - \underbrace{\nabla_{\mathbf{r}} E_F(\mathbf{r})}_{=0} + T(\mathbf{r})[E_c(\mathbf{r}) + E(\mathbf{k}) - E_F(\mathbf{r})]\underbrace{\nabla_{\mathbf{r}}(\frac{1}{T(\mathbf{r})})}_{=0} = 0, \tag{21.33}$$

gives us the constraints that $\mathbf{F} = -\nabla_{\mathbf{r}} E_c(\mathbf{r})$, and E_F and T cannot vary in space. Thus the general non-uniform equilibrium distribution form is indeed given by Equation 21.30. In Equation 21.33 the temperature cannot vary with $\mathbf{r}$ because $E(\mathbf{k})$ allows the choice of *any* $\mathbf{k}$, implying $\nabla_{\mathbf{r}}(1/T(\mathbf{r})) = 0$ must be met, and the rest of the terms must also individually be zero. Note that internal forces are allowed in equilibrium via $\mathbf{F} = -\nabla_{\mathbf{r}} E_c(\mathbf{r})$, with corresponding electric fields. The BTE is quantitatively enforcing what we knew all along, for example these are the internal electric fields in Schottky or pn junctions, or in heterostructures, as again indicated in Fig. 21.10. For electrons in the conduction band of 3D bulk silicon for example, the equilibrium distribution function is

$$f^0_{\mathbf{k}}(\mathbf{r}) = \frac{1}{\exp[\frac{E_c(\mathbf{r})+\frac{\hbar^2}{2}(\frac{k_x^2}{m_{xx}}+\frac{k_x^2}{m_{yy}}+\frac{k_x^2}{m_{zz}})-E_F}{k_b T}]+1}, \tag{21.34}$$

which explicitly shows the dependence on space via $\mathbf{r}$ and on the bandstructure via $\mathbf{k} = (k_x, k_y, k_z)$. If we consider a silicon pn junction in equilibrium, this expression holds everywhere: in the n-side, the p-side as well as inside the depletion region. Remembering that the BTE is defined separately for each band, the force $\mathbf{F}$ and the bandstructure of the valence band is defined along similar lines.

Non-uniform carrier distributions: The general non-uniform equilibrium electron concentration at any point $\mathbf{r}$ in a form suitable for non-uniform potentials and especially for quantized heterostructures is $n(\mathbf{r}) = \sum_{\mathbf{k}} |C_{\mathbf{k}}(\mathbf{r})|^2 f^0_{\mathbf{k}}(\mathbf{r})$, where $C_{\mathbf{k}}(\mathbf{r})$ is the $\mathbf{k}$-dependent **envelope function** of the wavepacket centered at $\mathbf{k}$. This result comes about from a mix of quantum mechanics and statistical mechanics in the following way. In the *absence* of scattering, the solution of the time-independent effective mass Schrödinger equation discussed in Chapters 14 and 16 is the time-independent envelope function[31] $C_{\mathbf{k}}(\mathbf{r})$. With a scattering potential $W(\mathbf{r}, t)$ the solution is a linear combination $\Psi(\mathbf{r}, t) = \sum_k \psi_{\mathbf{k}}(t) C_{\mathbf{k}}(\mathbf{r})$ which must satisfy

$$i\hbar \frac{\partial \Psi_{\mathbf{k}}(\mathbf{r}, t)}{\partial t} = [E_c(\mathbf{r}) + E(-i\nabla) + W(\mathbf{r}, t)]\Psi_{\mathbf{k}}(\mathbf{r}, t)$$
$$i\hbar \sum_k C_{\mathbf{k}}(\mathbf{r}) \frac{\partial \psi_{\mathbf{k}}(t)}{\partial t} = \sum_k \psi_{\mathbf{k}}(t)[E_c(\mathbf{r}) + E(\mathbf{k}) + W(\mathbf{r}, t)]C_{\mathbf{k}}(\mathbf{r}), \tag{21.35}$$

where multiplying by $C^{\star}_{\mathbf{k}'}(r)$, integrating over all space, and using the orthogonality $\int d\mathbf{r} C^{\star}_{\mathbf{k}'}(\mathbf{r}) C_{\mathbf{k}}(\mathbf{r}) = \delta_{\mathbf{k},\mathbf{k}'}$ gives

$$i\hbar \frac{\partial \psi_{\mathbf{k}}(t)}{\partial t} = \psi_{\mathbf{k}}(t)[E_c(\mathbf{r}) + E(\mathbf{k})] + \sum_{\mathbf{k}'} \psi_{\mathbf{k}'}(t) \langle C_{\mathbf{k}'}(\mathbf{r}) | W(\mathbf{r}, t) | C_{\mathbf{k}}(\mathbf{r}) \rangle$$
$$\implies \boxed{\frac{\partial \psi_{\mathbf{k}}(t)}{\partial t} + \frac{i}{\hbar}[E_c(\mathbf{r}) + E(\mathbf{k})]\psi_{\mathbf{k}}(t) = \sum_{\mathbf{k}'} \psi_{\mathbf{k}'}(t) \frac{W_{\mathbf{k}\mathbf{k}'}}{i\hbar}.} \tag{21.36}$$

Here $W_{\mathbf{k}\mathbf{k}'} = \langle C_{\mathbf{k}'}(\mathbf{r}) | W(\mathbf{r}, t) | C_{\mathbf{k}}(\mathbf{r}) \rangle = \int d\mathbf{r} C^{\star}_{\mathbf{k}'}(\mathbf{r}) W(\mathbf{r}, t) C_{\mathbf{k}}(\mathbf{r})$ is the same scattering matrix element that appears in Fermi's golden rule. Equation 21.36 shows how the time-dependent coefficients evolve in time due to scattering. In the absence of scattering $W(\mathbf{r}, t) = 0$ makes

[31] The actual wavefunction of the wavepacket is the product of the envelope and the time-independent periodic part of Bloch function $u_{\mathbf{k}}(\mathbf{r})$, this factor cancels out; indeed this is the reason why the effective mass approximation works in the first place. The spatial variation of the conduction band edge $E_c(\mathbf{r})$ is retained here in the WKB sense, discussed in Chapter 16, Section 16.4.

the RHS zero, and the amplitudes evolve in time in typical stationary-state fashion $\psi_k(t) \sim \exp[-i\frac{E_c(\mathbf{r})+E(\mathbf{k})}{\hbar}t]$. Statistical mechanics guarantees[32] that with scattering by *any* $W(\mathbf{r},t)$, the solution of Equation 21.36 satisfies $\langle\psi^\star_{\mathbf{k}'}(t)\psi_{\mathbf{k}}(t)\rangle = f^0_{\mathbf{k}}(\mathbf{r})\delta_{\mathbf{k},\mathbf{k}'}$ with $f^0_{\mathbf{k}}(\mathbf{r})$ given by Equation 21.30. Then, the electron distribution at equilibrium is given by

$$n(\mathbf{r},t) = \Psi^\star(\mathbf{r},t)\Psi(\mathbf{r},t) = \sum_{\mathbf{k},\mathbf{k}'} \underbrace{\psi^\star_{\mathbf{k}'}(t)\psi_{\mathbf{k}}(t)}_{\langle\psi^\star_{\mathbf{k}'}(t)\psi_{\mathbf{k}}(t)\rangle = f^0_{\mathbf{k}}(\mathbf{r})\delta_{\mathbf{k},\mathbf{k}'}} \quad C^\star_{\mathbf{k}'}(\mathbf{r})C_{\mathbf{k}}(\mathbf{r})$$

$$\implies \boxed{n_0(\mathbf{r}) = \sum_{\mathbf{k}} |C_{\mathbf{k}}(\mathbf{r})|^2 f^0_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{k}} \frac{|C_{\mathbf{k}}(\mathbf{r})|^2}{\exp[\frac{E_c(\mathbf{r})+E(\mathbf{k})-E_F}{k_bT}]+1}.} \tag{21.37}$$

As an example of its application, consider an infinitely deep quantum well in the z-direction shown in Fig. 21.11, where the electron is free to move in the $x-y$ plane. The electron envelope function from Chapter 14 is given by $C_{n_z}(x,y,z) = [\sqrt{2/L_z}\sin(\pi n_z z/L_z)] \cdot e^{i(k_x x+k_y y)}/\sqrt{L_x L_y}$ and eigenvalues $E_{n_z} + \hbar^2(k_x^2+k_y^2)/(2m_c^\star)$. If the Fermi level is at E_F the electron distribution is given by

$$n_0(x,y,z) = g_s g_v \sum_{k_x,k_y,n_z} \frac{|C_{n_z}(x,y,z)|^2}{\exp[\frac{E_{n_z}+\frac{\hbar^2(k_x^2+k_y^2)}{2m_c^\star}-E_F}{k_bT}]+1}$$

$$= \frac{2g_s g_v}{L_z}\sum_{n_z}\sin^2(\frac{\pi n_z z}{L_z}) \cdot \frac{1}{(2\pi)^2}\int_{k=0}^{\infty}\frac{2\pi k dk}{\exp[\frac{\frac{\hbar^2k^2}{2m_c^\star}-(E_F-E_{n_z})}{k_bT}]+1}$$

$$\implies n_0(z) = \frac{g_s g_v m_c^\star k_b T}{\pi\hbar^2 L_z}\sum_{n_z}\sin^2(\frac{\pi n_z z}{L_z}) \cdot \ln[1+e^{\frac{E_F-E_{n_z}}{k_bT}}]. \tag{21.38}$$

21.6 The RTA: time to relax!

We resume our journey towards the promised land of actually solving the BTE to obtain the $f_{\mathbf{k}}(\mathbf{r},t)$ in response to *external* Lorentz forces due to electric and magnetic fields $\mathbf{F} = -q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$, or temperature gradients $\nabla_{\mathbf{r}}[\frac{1}{T(\mathbf{r})}]$. The BTE with such forces is

$$\frac{\partial f_{\mathbf{k}}}{\partial t} + \underbrace{\mathbf{v}\cdot\nabla_{\mathbf{r}} f_{\mathbf{k}} + \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_{\mathbf{k}}}_{\hat{\mathcal{L}}\{f_{\mathbf{k}}\}} = \hat{C}\{f_{\mathbf{k}}\}, \tag{21.39}$$

where the force term is labeled as $\hat{\mathcal{L}}\{f_{\mathbf{k}}\}$. For general inelastic scattering, from Equation 21.29, $\hat{C}\{f_{\mathbf{k}}\} = \sum_{\mathbf{k}'} S(\mathbf{k}\to\mathbf{k}')\times$ $[\exp[(E_{\mathbf{k}'}-E_{\mathbf{k}})/k_bT]f_{\mathbf{k}'}(1-f_{\mathbf{k}}) - f_{\mathbf{k}}(1-f_{\mathbf{k}'})]$ makes the RHS non-linear in $f_{\mathbf{k}}$. The sum over $\mathbf{k}'$ converts to an integral, making the BTE in its full glory a non-linear integro–differential equation, and in several dimensions to boot, which is difficult to solve. But a remarkable simplification is possible by appealing to the physics for an approximation.

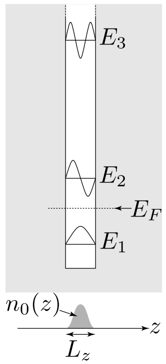


Fig. 21.11 Equilibrium electron distribution in an infinitely deep quantum well. Highly non-uniform electron distributions in quantized structures are obtained by using the effective mass approximation with the generalized equilibrium Fermi–Dirac distribution function using Equation 21.37. Note that the distribution $n_0(\mathbf{r})$ here is only the envelope portion; the full wavefunction will also have a Bloch-periodic part with the periodicity of the lattice, of the form $n_0(\mathbf{r})|u(\mathbf{r})|^2$. In quantum well HEMTs and lasers, the E_F is tuned by field effect or carrier injection, thereby controlling the operation of such devices.

---

[32] Though we do not discuss the rigorous proof, it is justified in the following heuristic way. The Schrödinger equation with scattering may be solved in principle (if not in practice) at any instant t with the scattering potential $W(\mathbf{r},t)$ to reveal all eigenfunctions at that instant, which by the expansion theorem are the linear combinations $\sum_{\mathbf{k}}\psi_{\mathbf{k}}(t)C_{\mathbf{k}}(\mathbf{r})$. From statistical mechanics we know the *average* equilibrium occupation *probability* of state $\mathbf{k}$ is $f^0_{\mathbf{k}}(\mathbf{r})$, which must be related to the quantum mechanical *amplitude* via $|\psi_{\mathbf{k}}(t)|^2$. But because of scattering by the random potential $W(\mathbf{r},t)$, electrons transition between states $\mathbf{k}\leftrightarrow\mathbf{k}'$, their amplitudes are mixed, causing fluctuations, resulting in $\langle\psi^\star_{\mathbf{k}}(t)\psi_{\mathbf{k}'}(t)\rangle = f^0_{\mathbf{k}}(\mathbf{r})\delta_{\mathbf{k},\mathbf{k}'}$. $W(\mathbf{r})$ that are independent of time cause elastic scattering that does not randomize the phase of $\psi_{\mathbf{k}}(t)$; only the amplitudes fluctuate. But time-dependent $W(\mathbf{r},t)$ cause inelastic scattering, and the phases also fluctuate. Fluctuations are measurable as macroscopic noise in electrical currents and voltages and reveal information about the microscopic $W(\mathbf{r},t)$.

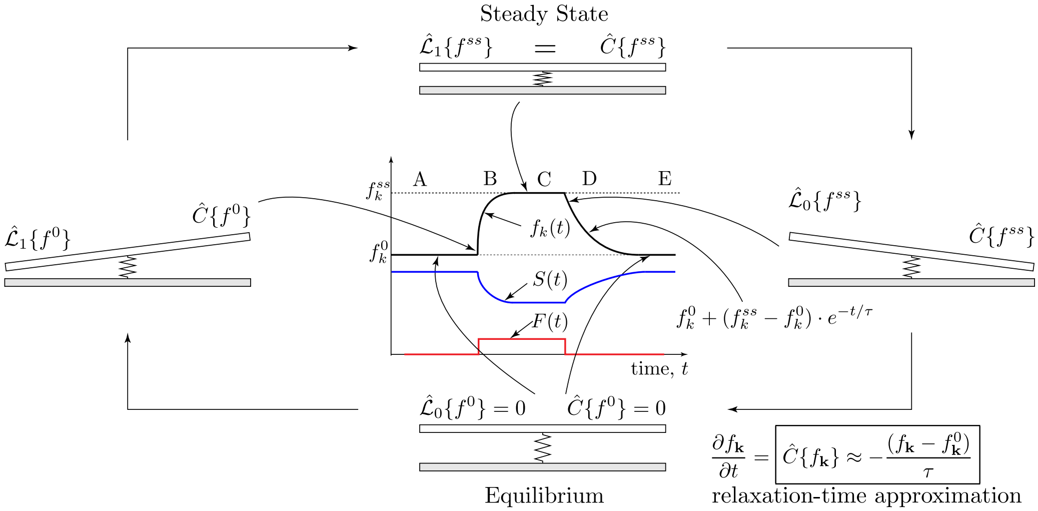

Steady State
$\hat{\mathcal{L}}_1\{f^{ss}\} = \hat{C}\{f^{ss}\}$
$\hat{C}\{f^0\}$
$\hat{\mathcal{L}}_1\{f^0\}$
f_k^{ss}
f_k^0
A
B
C
D
E
$f_k(t)$
$S(t)$
$F(t)$
time, t
$\hat{\mathcal{L}}_0\{f^{ss}\}$
$\hat{C}\{f^{ss}\}$
$f_k^0 + (f_k^{ss} - f_k^0) \cdot e^{-t/\tau}$
$\hat{\mathcal{L}}_0\{f^0\} = 0$
$\hat{C}\{f^0\} = 0$
Equilibrium
$\frac{\partial f_{\mathbf{k}}}{\partial t} = \boxed{\hat{C}\{f_{\mathbf{k}}\} \approx -\frac{(f_{\mathbf{k}} - f_{\mathbf{k}}^0)}{\tau}}$
relaxation-time approximation

Fig. 21.12 Schematic depiction of the balance between the LHS force term $\hat{\mathcal{L}}\{f\}$ and the collision term $\hat{C}\{f\}$ of the BTE when the system is taken out of, and then returned to equilibrium. The center figure shows the evolution of the distribution function $f_k(t)$, the external force $F(t)$, and the entropy $S(t)$ with time. Fig. 21.13 shows the time evolution in **k**-space.

Consider Fig. 21.12, in which the time evolution of the electron distribution function $f_{\mathbf{k}}(t)$, the force $F(t)$, and the corresponding entropy $S(t)$ of the electron system are shown schematically[33]. At equilibrium, there are no external forces and the entropy is at its maximum. In the LHS $\hat{\mathcal{L}}_0 = 0$, the RHS $\hat{C}\{f_{\mathbf{k}}^0\} = 0$, and they "balance", indicated by the see-saw at the bottom. The instant an external force is turned on, $\hat{\mathcal{L}}_0 \to \hat{\mathcal{L}}_1 \neq 0$, and the balance is broken because the LHS is non-zero while the RHS is zero. The entropy decreases, and $f_{\mathbf{k}}(t)$ will now change to find a new steady-state value $f_{\mathbf{k}}^{ss}$ which restores the balance in the BTE as shown on the top. If after steady-state is reached, the external force is removed, the electron distribution must relax back from $f_{\mathbf{k}}^{ss} \to f_{\mathbf{k}}^0$. If $f_{\mathbf{k}}$ did not deviate too far from $f_{\mathbf{k}}^0$, when returning to equilibrium it is fair to approximate the BTE by[34]

$$\frac{\partial f_{\mathbf{k}}}{\partial t} = \boxed{\hat{C}\{f_{\mathbf{k}}\} \approx -\frac{(f_{\mathbf{k}} - f_{\mathbf{k}}^0)}{\tau}} \implies f_{\mathbf{k}}(t) \approx f_{\mathbf{k}}^0 + (f_{\mathbf{k}}^{ss} - f_{\mathbf{k}}^0)e^{-\frac{t}{\tau}}. \tag{21.40}$$

If the BTE can be written in such a linear fashion, its solution in Equation 21.40 indicates that $f_k(t)$ will relax from f_k^{ss} to f_k^0 as $t \to \infty$ exponentially with a characteristic relaxation time τ. This is the **relaxation time approximation (RTA)** of the BTE. Its simplicity underpins its immense practical value. From the discussion of equilibrium, all microscopic scattering processes are captured by the time τ.

The path out of equilibrium with this approximation is

$$\frac{\partial f_{\mathbf{k}}}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{r}} f_{\mathbf{k}} + \frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f_{\mathbf{k}} = \hat{C}\{f_{\mathbf{k}}\} \approx -\frac{(f_{\mathbf{k}} - f_{\mathbf{k}}^0)}{\tau}. \tag{21.41}$$

[33] The only known exact solution of the BTE is for a specialized scattering potential for the classical ideal gas, obtained by James Clerk Maxwell around the time Boltzmann formulated the equation.

[34] This is also justified through a series solution of the form $f_{\mathbf{k}} \approx f_{\mathbf{k}}^0 + f_{\mathbf{k}}^{\{1\}} + f_{\mathbf{k}}^{\{2\}} + ...$, where $f_{\mathbf{k}}^{\{1\}}$ is the first order, $f_{\mathbf{k}}^{\{2\}}$ is the second order, and so on. If the second and higher orders are neglected, the first order solution is indeed in the form $f_{\mathbf{k}}^{\{1\}} \approx (f_{\mathbf{k}} - f_{\mathbf{k}}^0)$, justifying the relaxation-time approximation.

The simplified solution was discussed in Chapter 16, Section 16.6. For a force **F** and concentration gradient $\nabla_{\mathbf{r}} f_{\mathbf{k}}$, when the distribution approaches steady state ($\partial f_{\mathbf{k}}/\partial t = 0$), the solution to this equation is

$$\boxed{f_{\mathbf{k}} = f_{\mathbf{k}}^{ss} \approx f_{\mathbf{k}}^{0} + \tau \mathbf{v}_{\mathbf{k}} \cdot [\underbrace{(-\frac{\partial f_{\mathbf{k}}^{0}}{\partial E_{\mathbf{k}}})\mathbf{F}}_{\text{drift}} + \underbrace{(-\nabla_{\mathbf{r}} f_{\mathbf{k}}^{0})}_{\text{diffusion}}],} \tag{21.42}$$

where the group velocity $\mathbf{v}_{\mathbf{k}} = \hbar^{-1}\nabla_{\mathbf{k}} E_{\mathbf{k}}$ and the small deviation approximation $-\partial f_{\mathbf{k}}/\partial E_{\mathbf{k}} \approx -\partial f_{\mathbf{k}}^{0}/\partial E_{\mathbf{k}}$ are used. The two components indicated underpin the very useful **drift-diffusion** method of analysis of charge transport in semiconductors. Fig. 21.13 indicates the time evolution of the distribution function, depicting possible $f_{\mathbf{k}}$ in a 2D system at the various instants of time A–E indicated in Fig. 21.12. The solution in Equation 21.42 gives $f_{\mathbf{k}}^{ss}$, the steady state distribution at time instant C. The shift of the occupation function depends roughly linearly on τ, leading to a higher conductivity for larger τ.

This approximate solution of the BTE expresses the unknown distribution function $f_{\mathbf{k}}$ in terms of all known quantities and driving forces on the RHS. Several transport parameters can now be evaluated from it. Consider the flow of quantities such as charge, heat, spin, etc that depend on the wavevector **k** through a known functional form $\phi_{\mathbf{k}}$ with a corresponding volume density $\rho_{\phi} = f_{\mathbf{k}}\phi_{\mathbf{k}}/L^{d}$. The net ϕ-current $\mathbf{J}_{\phi}$, and a corresponding ϕ-conservation rule from Equation 21.42 are

$$\boxed{\mathbf{J}_{\phi} = \frac{1}{L^{d}}\sum_{\mathbf{k}} \phi_{\mathbf{k}} \mathbf{v}_{\mathbf{k}} f_{\mathbf{k}}} \text{ and } \boxed{\frac{\partial \rho_{\phi}}{\partial t} + \nabla_{\mathbf{r}} \cdot \mathbf{J}_{\phi} = G_{\phi} - R_{\phi},} \tag{21.43}$$

where G_{ϕ}, R_{ϕ} are generation and recombination rates and L^{d} is the volume in d-dimensions. We use the current and continuity expressions in the following sections to understand the effect of scattering on electrical conductivity and general transport coefficients of semiconductors. In applying the approximate solution of Equation 21.42 in evaluating various currents, we must exercise caution because as discussed in the last section, a relaxation time may be defined for elastic scattering but not for inelastic scattering. When defined, τ is in general dependent on **k**, as was seen in the preview of ionized impurity scattering in Chapter 16, Section 16.8. We will discuss its validity as we use it to connect to several experimental transport parameters.

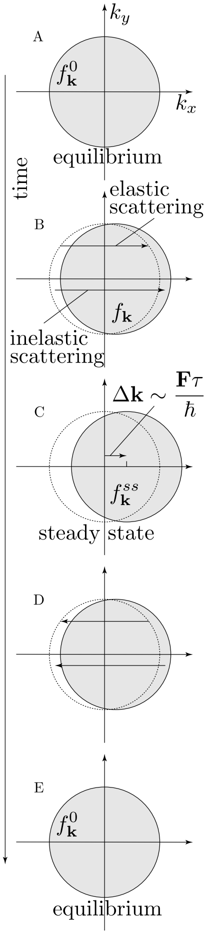


Fig. 21.13 Schematic diagram showing the distribution function $f_{\mathbf{k}}(t)$ of Fig. 21.12 in **k**-space at times indicated by A, B, C, D, and E. A circular Fermi-surface results from an isotropic effective mass. Note the elastic and inelastic scattering processes indicated. The Fermi-circle does not roll away, but gets "stuck" at a distance roughly $\Delta k \sim \mathbf{F}\tau/\hbar$ from the origin for a force **F**. When the force is removed, the distribution function returns to equilibrium. The non-equilibrium $f_{\mathbf{k}}$ shape does not need to remain circular in the general case, even for isotropic effective masses. A typical relaxation time of $\tau \sim 10^{-13}$s, or 0.1 ps at room temperature for electron-phonon scattering in semiconductors and a small electric field 1 kV/cm leads to $|\Delta\mathbf{k}| \sim 10^{7}$/m, less than $(1/100)^{\text{th}}$ of the Brillouin-zone edge.

21.7 One formula to rule them all

Fig. 21.14 shows how the entropy of the electron system changes when external forces are applied. Recall that the entropy of state **k** of the electron system with distribution function $f_{\mathbf{k}}$ is $s_{\mathbf{k}} = -k_b[f_{\mathbf{k}}\ln(f_{\mathbf{k}}) + (1-f_{\mathbf{k}})\ln(1-f_{\mathbf{k}})]$, and its total entropy is $\mathcal{S} = \sum_{\mathbf{k}} s_{\mathbf{k}}$. The three quantities $f_{\mathbf{k}}$, $s_{\mathbf{k}}$, and $\mathcal{S}$ are shown in equilibrium and in steady state. Fig. 21.14 (a) shows that an external force such as an electric field changes the distribution function from $f^0_{\mathbf{k}} \to f_{\mathbf{k}} = f^0_{\mathbf{k}} - g_{\mathbf{k}}(\partial f^0_{\mathbf{k}}/\partial E_{\mathbf{k}})$, where $g_{\mathbf{k}} = -\tau \mathbf{F} \cdot \mathbf{v}_{\mathbf{k}}$ is a perturbation with energy units, characterized by the relaxation time τ, the external force $\mathbf{F}$, and the electron bandstructure via group velocity $\mathbf{v}_{\mathbf{k}}$. The distribution function reaches a steady state $f^{ss}_{\mathbf{k}}$ if the external force is held constant (the dynamics was illustrated in Figures 21.12 and 21.13). Fig.21.14 (b) shows that as a result, the entropy density $s_{\mathbf{k}}$ changes from its equilibrium value $s^0_{\mathbf{k}}$ to the steady-state value $s^{ss}_{\mathbf{k}}$. Fig. 21.14 (c) shows the total entropy $\mathcal{S} = \sum_{\mathbf{k}} s_{\mathbf{k}}$ as a function of the external perturbation g, indicating that the **entropy of the electron system decreases when an external force is applied**. This is because an electron distribution drifting in a direction has more order, and hence less entropy. When the external force is removed, $f^{ss}_{\mathbf{k}} \to f^0_{\mathbf{k}}$, $s^{ss}_{\mathbf{k}} \to s^0_{\mathbf{k}}$, and $\mathcal{S}^{ss} \to \mathcal{S}^0$ in a characteristic time of τ. The total entropy at equilibrium S^0 is its maximum value.

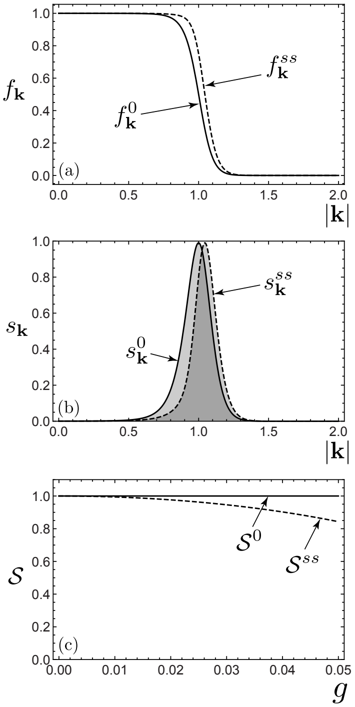


Fig. 21.14 The distribution function $f_{\mathbf{k}}$ in (a) and entropy density $s_{\mathbf{k}}$ in (b) shown as a function of $k = |\mathbf{k}|$ in equilibrium (solid lines), and steady state (dashed line). The area under $s_{\mathbf{k}}$ is the total entropy of the electron system, shown in (c) as a function of the external perturbation $g = g_{\mathbf{k}}$. The entropy of the electron system decreases compared to its equilibrium value when an external force is applied. Notice that the entropy density is zero in regions where $f_{\mathbf{k}} = 0$ or $f_{\mathbf{k}} = 1$, and peaks in the transition region. Also note that the area under the entropy density curve $s^{ss}_{\mathbf{k}}$ decreases compared the equilibrium density curve $s^0_{\mathbf{k}}$.

We will now prove that **the rate of change of entropy due to collisions and scattering is always positive**: $d\mathcal{S}_{\text{coll.}}/dt \geq 0$, **and that due to external forces is negative**: $d\mathcal{S}_{\text{drift}}/dt \leq 0$. Taking the time derivative of $\mathcal{S} = -k_b \sum_{\mathbf{k}}[f_{\mathbf{k}} \ln(f_{\mathbf{k}}) + (1 - f_{\mathbf{k}}) \ln(1 - f_{\mathbf{k}})]$, using $f_{\mathbf{k}} = f^0_{\mathbf{k}} - g_{\mathbf{k}}(\partial f^0_{\mathbf{k}}/\partial E_{\mathbf{k}})$, and retaining up to linear terms in $g_{\mathbf{k}}$ gives

$$\frac{d\mathcal{S}}{dt} = k_b \sum_{\mathbf{k}} \frac{df_{\mathbf{k}}}{dt} \ln(\frac{f_{\mathbf{k}}}{1-f_{\mathbf{k}}}) \approx -k_b \sum_{\mathbf{k}} \frac{df_{\mathbf{k}}}{dt}[\frac{g_{\mathbf{k}}}{k_b T} + \frac{E-\mu}{k_b T}] \approx -\frac{1}{T}\sum_{\mathbf{k}} g_{\mathbf{k}} \frac{df_{\mathbf{k}}}{dt}, \tag{21.44}$$

where the term with $(E - \mu)/k_b T$ leads to a net energy increase and non-ohmic effects, and is neglected. Writing $\mathcal{S} = \mathcal{S}_{\text{coll.}} + \mathcal{S}_{\text{drift}}$ we get

$$\frac{d\mathcal{S}_{\text{coll.}}}{dt} = -\frac{1}{T}\sum_{\mathbf{k}} g_{\mathbf{k}} \frac{df_{\mathbf{k}}}{dt}|_{\text{coll.}} \text{ and } \frac{d\mathcal{S}_{\text{drift}}}{dt} = -\frac{1}{T}\sum_{\mathbf{k}} g_{\mathbf{k}} \frac{df_{\mathbf{k}}}{dt}|_{\text{drift}}. \tag{21.45}$$

Using $f_{\mathbf{k}} = f^0_{\mathbf{k}} - g_{\mathbf{k}} \frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}$ in the collision term in the RHS of the BTE[35],

$$\begin{aligned}
\frac{df_{\mathbf{k}}}{dt}|_{\text{coll.}} &= \sum_{\mathbf{k}'}[S(\mathbf{k}' \to \mathbf{k}) f_{\mathbf{k}'}(1 - f_{\mathbf{k}}) - S(\mathbf{k} \to \mathbf{k}') f_{\mathbf{k}}(1 - f_{\mathbf{k}'})] \\
&= \sum_{\mathbf{k}'}[S(\mathbf{k} \to \mathbf{k}') f^0_{\mathbf{k}}(1 - f^0_{\mathbf{k}'})[\frac{f_{\mathbf{k}'}(1 - f_{\mathbf{k}})}{f^0_{\mathbf{k}'}(1 - f^0_{\mathbf{k}})} - \frac{f_{\mathbf{k}}(1 - f_{\mathbf{k}'})}{f^0_{\mathbf{k}}(1 - f^0_{\mathbf{k}'})}] \\
\implies \frac{df_{\mathbf{k}}}{dt}|_{\text{coll.}} &= \frac{1}{k_b T} \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}') f^0_{\mathbf{k}}(1 - f^0_{\mathbf{k}'})[g_{\mathbf{k}'} - g_{\mathbf{k}}],
\end{aligned} \tag{21.46}$$

where in the last step we have retained terms to first order in $g_{\mathbf{k}}$. Using this form of $\frac{df_{\mathbf{k}}}{dt}|_{\text{coll.}}$ in Equation 21.45 for collision gives

---

[35] Here we use the microscopic reversibility relation $S(\mathbf{k}' \to \mathbf{k}) f^0_{\mathbf{k}'}(1 - f^0_{\mathbf{k}}) = S(\mathbf{k} \to \mathbf{k}') f^0_{\mathbf{k}}(1 - f^0_{\mathbf{k}'})$, and the identity $\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}} = -\frac{f^0_{\mathbf{k}}(1 - f^0_{\mathbf{k}})}{k_b T}$.

$$\begin{aligned} \frac{d\mathcal{S}_{\text{coll.}}}{dt} &= \frac{1}{k_b T^2}\sum_{\mathbf{k},\mathbf{k}'} S(\mathbf{k}\to\mathbf{k}') f^0_{\mathbf{k}}(1-f^0_{\mathbf{k}'}) g_{\mathbf{k}}[g_{\mathbf{k}} - g_{\mathbf{k}'}] \\ g_{\mathbf{k}} \leftrightarrow g_{\mathbf{k}'} \implies &= \frac{1}{k_b T^2}\sum_{\mathbf{k},\mathbf{k}'} S(\mathbf{k}\to\mathbf{k}') f^0_{\mathbf{k}}(1-f^0_{\mathbf{k}'}) g_{\mathbf{k}'}[g_{\mathbf{k}'} - g_{\mathbf{k}}] \\ \implies \frac{d\mathcal{S}_{\text{coll.}}}{dt} &= \frac{1}{2k_b T^2}\sum_{\mathbf{k},\mathbf{k}'} S(\mathbf{k}\to\mathbf{k}') f^0_{\mathbf{k}}(1-f^0_{\mathbf{k}'}) [g_{\mathbf{k}} - g_{\mathbf{k}'}]^2 \geq 0, \end{aligned} \tag{21.47}$$

where we have used the fact that the sum over $\mathbf{k}, \mathbf{k}'$ cannot change upon interchanging $g_{\mathbf{k}} \leftrightarrow g_{\mathbf{k}'}$, and then added the first two equations and divided by 2 to get the third. Since all terms in the RHS of the third form of $\frac{d\mathcal{S}_{\text{coll.}}}{dt}$ are positive, including the perturbation which appears as $[g_{\mathbf{k}} - g_{\mathbf{k}'}]^2$, we have proven that the rate of production of entropy due to collisions is always positive[36].

To prove $\frac{d\mathcal{S}_{\text{drift}}}{dt} \leq 0$, we first note that when the electron distribution is in the steady-state, the net change of entropy is zero, implying $\frac{d\mathcal{S}}{dt} = \frac{d(\mathcal{S}_{\text{coll.}}+\mathcal{S}_{\text{drift}})}{dt} = 0$, and since $\frac{d\mathcal{S}_{\text{coll.}}}{dt} \geq 0$, it follows that at steady state, $\frac{d\mathcal{S}_{\text{drift}}}{dt} \leq 0$. The proof for the general case provides more insight into the physics. The LHS drift term of the BTE with Equation 21.45 gives[37]

$$\begin{aligned} \frac{df_{\mathbf{k}}}{dt}|_{\text{drift}} &= \frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}} f_{\mathbf{k}} + \mathbf{v}_{\mathbf{k}}\cdot\nabla_{\mathbf{r}} f_{\mathbf{k}} \\ &\approx -\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}\mathbf{v}_{\mathbf{k}}\cdot[\mathbf{F} + \nabla_{\mathbf{r}}(E_{\mathbf{k}}-\mu) + (E_{\mathbf{k}}-\mu)T\nabla_{\mathbf{r}}(\frac{1}{T})] \\ \frac{d\mathcal{S}_{\text{drift}}}{dt} &= -\frac{1}{T}\sum_{\mathbf{k}} \underbrace{g_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})}_{f_{\mathbf{k}}-f^0_{\mathbf{k}}}\mathbf{v}_{\mathbf{k}}\cdot[\mathbf{F} + \nabla_{\mathbf{r}}(E_{\mathbf{k}}-\mu) + (E_{\mathbf{k}}-\mu)T\nabla_{\mathbf{r}}(\frac{1}{T})] \\ &= -\frac{1}{T}(\mathbf{F}-\nabla_{\mathbf{r}}\mu)\cdot\underbrace{\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}} f_{\mathbf{k}}}_{L^d\cdot\mathbf{j}} - \nabla_{\mathbf{r}}(\frac{1}{T})\cdot\underbrace{\sum_{\mathbf{k}}(E_{\mathbf{k}}-\mu)\mathbf{v}_{\mathbf{k}} f_{\mathbf{k}}}_{L^d\cdot(\mathbf{J}_u-\mu\mathbf{J})} \\ \implies \frac{d\mathcal{S}_{\text{drift}}}{dt} &= -\frac{L^d}{T}[\underbrace{(\mathbf{F}-\nabla_{\mathbf{r}}\mu)\cdot\mathbf{j}}_{\text{Joule heat}} + T\nabla_{\mathbf{r}}(\frac{1}{T})\cdot(\underbrace{\mathbf{J}_u-\mu\mathbf{J}}_{\text{heat current}})] \leq 0. \end{aligned} \tag{21.48}$$

Several important currents are thus defined using entropy[38]. They are: **particle current** $\mathbf{j} = (\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}} f_{\mathbf{k}})/L^d$, **charge current** $\mathbf{J} = (-q)\mathbf{j}$, **energy current** $\mathbf{J}_u = (\sum_{\mathbf{k}} E_{\mathbf{k}}\mathbf{v}_{\mathbf{k}} f_{\mathbf{k}})/L^d$, and **heat current** $\mathbf{J}_q = \mathbf{J}_u - \mu\mathbf{J}$. The rate of entropy production due to the drift term is thus negative, because both terms in the square bracket are positive. The reason is very physical: the driving force $\mathbf{F} - \nabla_{\mathbf{r}}\mu$ in the first term is the net electrochemical force. The first term is this force times the resulting current $\mathbf{J}$,

[36] Because collisions randomize the velocities and energies, $\frac{d\mathcal{S}_{\text{coll.}}}{dt} \geq 0$ makes physical sense. Note that $\frac{d\mathcal{S}_{\text{coll.}}}{dt} \geq 0$ is seen in the slope of $\mathcal{S}(t)$ during the relaxation phase labeled "D" in Fig. 21.12. The balancing drift portion where $\frac{d\mathcal{S}_{\text{drift}}}{dt} \leq 0$ is labeled 'B' in the same figure. Both are shown in the $\mathbf{k}$-space in Fig. 21.13. We recognize this behavior as the Boltzmann H-theorem again, but now for the particular solution $f_{\mathbf{k}} = f^0_{\mathbf{k}} - g_{\mathbf{k}}(\partial f^0_{\mathbf{k}}/\partial E_{\mathbf{k}})$ of the BTE.

[37] For a semiconductor, $E_{\mathbf{k}} - \mu = E_{\mathbf{k}} - (E_F - E_c) = (E_c + E_{\mathbf{k}}) - E_F$, the argument of the exponent in the Fermi–Dirac function. Note that $\nabla_{\mathbf{r}} E_{\mathbf{k}} = 0$, and so is $\sum_{\mathbf{k}} f^0_{\mathbf{k}} v_{\mathbf{k}}$ because $f^0_{\mathbf{k}}$ is symmetrical in $\mathbf{k}$ but $\mathbf{v}_{\mathbf{k}}$ is antisymmetrical.

[38] The notation L^d implies a volume in d-dimensions. Note that at equilibrium, all the defined net currents are zero.

and since $\mathbf{J} \propto \mathbf{F} - \nabla_{\mathbf{r}}\mu$, the product must be positive. This term is the Joule heat produced by the drifting electrons. The driving force in the second term $\nabla_{\mathbf{r}}(\frac{1}{T})$ is a gradient in the temperature (or inverse temperature). Since the resulting heat current carried by electrons flows from higher to lower temperature, the product again must be positive.

The proof of the negative sign of the rate of drift entropy change gives us a first glimpse of a deeper principle[39]: that **the rate of entropy production can be written as a sum of products of driving forces and resulting currents**. It will pay off handsomely to develop the concept further. We will see now from thermodynamic arguments why **the flow of electrons and heat is required to ensure the conservation of particles and energy, and the direction of their flow is dictated by entropy**. This consideration will reveal a richer range of transport phenomena, in which the charge current is only a part of the puzzle, with heat current playing an important role. There is also a deep connection between them, first uncovered by Onsager.

The first law of thermodynamics states that the energy dU added to a system partitions into doing work PdV, adding particles μdN, and causing irrecoverable entropy $Td\mathcal{S}$ via the energy conservation relation $dU = Td\mathcal{S} + pdV + \mu dN$. For the electron system in solids, $dV = 0$. The volume densities $u = \frac{U}{V}$, $s = \frac{\mathcal{S}}{V}$, and $n = \frac{N}{V}$ give the time dependence

$$dU = TdS + \mu dN \implies \frac{\partial u}{\partial t} = T\frac{\partial s}{\partial t} + \mu\frac{\partial n}{\partial t}. \tag{21.49}$$

Because the particle density n and energy density u follow conservation relations, inserting them into Equation 21.49 gives a continuity equation and current density $\mathbf{J}_s = \frac{\mathbf{J}_u - \mu\mathbf{J}}{T}$ for the entropy[40]:

$$\begin{aligned}
\frac{\partial n}{\partial t} + \nabla_{\mathbf{r}}\cdot\mathbf{j} = 0 \text{ and } \frac{\partial u}{\partial t} + \nabla_{\mathbf{r}}\cdot\mathbf{J}_u = \mathbf{F}\cdot\mathbf{J} \implies \\
\frac{\partial s}{\partial t} = \frac{1}{T}(\frac{\partial u}{\partial t} - \mu\frac{\partial n}{\partial t}) = \frac{1}{T}[-\nabla_{\mathbf{r}}\cdot(\mathbf{J}_u - \mu\mathbf{J}) + \mathbf{F}\cdot\mathbf{J}] \\
\frac{\partial s}{\partial t} + \nabla_{\mathbf{r}}\cdot(\underbrace{\frac{\mathbf{J}_u - \mu\mathbf{J}}{T}}_{\mathbf{J}_s}) = \frac{1}{T}[\underbrace{(-q\mathbf{j})}_{\mathbf{J}}\cdot(-q\underbrace{\nabla_{\mathbf{r}}(\frac{\mu}{q} - \varphi))}_{\mathbf{E}} + \underbrace{(\mathbf{J}_u - \mu\mathbf{J})}_{\mathbf{J}_q}\cdot\underbrace{(\frac{-\nabla_{\mathbf{r}}T}{T})}_{\mathbf{F}_q}] \\
\implies \boxed{\frac{\partial s}{\partial t} + \nabla_{\mathbf{r}}\cdot\mathbf{J}_s = \sum_i \mathbf{J}_i\cdot\mathbf{F}_i,}
\end{aligned} \tag{21.50}$$

where the generation rate of entropy in the RHS is expressed as a sum of products of forces $\mathbf{F}_i$ and currents $\mathbf{J}_i$. A **theorem of irreversible thermodynamics** states that each current depends on *every* force[41]:

$$\frac{\partial\mathcal{S}}{\partial t} = \sum_i J_i F_i \implies J_i = \sum_j L_{ij}F_j \text{ by matrix inversion.} \tag{21.51}$$

The theorem relates the currents that flow in response to the force via linear transport coefficients L_{ij}. Using this theorem, we now formulate

[39] The discussion here attempts to answer the question: "*Why* does transport of electrons or heat occur"? The dynamical laws of quantum (or classical) mechanics can answer *what* processes occur, but not *why*, because the dynamical laws are symmetrical in time, and are therefore reversible. The answer to *why* is provided by thermodynamics because it identifies an arrow of time and explains irreversibility. It is explicitly seen that current conduction is a direct consequence of the production of entropy.

[40] $\mathbf{F}\cdot\mathbf{J}$ is the conventional Joule heat.

[41] Here is the meaning of the theorem: if across a semiconductor there is both a gradient in electric potential and a gradient in temperature, then a charge current will flow due to both gradients, and a heat current will also flow due to both gradients. The charge and heat currents, and hence their coefficients, are linked.

a generalized expression for transport coefficients (Equation 21.55) in response to forces that arise from gradients in the electrochemical potential and temperature, a "formula that rules them all" so to speak.

To find a generalized solution of the BTE in the RTA of the form $f_{\mathbf{k}} = f^0_{\mathbf{k}} - \frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}} g_{\mathbf{k}}$ in the *simultaneous* presence of an electric force $\mathbf{F}$, a magnetic field $\mathbf{B}$, and a temperature gradient $\nabla_{\mathbf{r}} T$, let's use the trial form $g_{\mathbf{k}} = -\tau_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}} \cdot \mathbf{G})$, where the vector $\mathbf{G}$ is a generalized electro-magneto-thermal force. The electro-thermal forces, being *polar vectors*, are clubbed into[42] $\mathbf{F} = \nabla_{\mathbf{r}}(\mu - q\varphi) + \frac{E-\mu}{T}\nabla_{\mathbf{r}} T$. The magnetic field, being an *axial vector*, is given special status. By substituting this trial form into the BTE, making a small perturbation approximation, and using $\mu = q\tau/m^\star$, a generalized force vector $\mathbf{G}$ is obtained:

$$\mathbf{G} = \frac{1}{(1+\mu^2|\mathbf{B}|^2)}[\mathbf{F} + \mu(\mathbf{F}\times\mathbf{B}) + \mu^2(\mathbf{F}\cdot\mathbf{B})\mathbf{B}]. \tag{21.52}$$

The $\mathbf{B} = 0$ solution is $\mathbf{G} = \mathbf{F}$, an extension of Equation 21.42, but now including the temperature gradient explicitly. For $\mathbf{B} \neq 0$ the solution is much richer[43]. This solution of the BTE immediately yields the particle current $\mathbf{j}$ in response to $\mathbf{E} = \mathbf{F}/Q$, $\mathbf{B}$, and $\nabla_{\mathbf{r}} T$ as the sum

$$\begin{aligned}
\mathbf{j} &= \frac{1}{L^d}\sum_{\mathbf{k}} \mathbf{v}_{\mathbf{k}}\left(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}\right) g_{\mathbf{k}} = \frac{1}{L^d}\sum_{\mathbf{k}} \mathbf{v}_{\mathbf{k}}\left(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}\right)(-\tau_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}\cdot\mathbf{G}), \\
\implies \mathbf{J}/Q = \mathbf{j} &= [\underbrace{K_{10}\mathbf{F}}_{\propto\mathbf{E}} + \underbrace{K_{20}(\mathbf{B}\times\mathbf{F})}_{\text{Hall effect}} + \underbrace{K_{30}\mathbf{B}(\mathbf{B}\cdot\mathbf{F})}_{\text{magnetoresistance}}] \\
&+[\underbrace{K_{11}\frac{\nabla_{\mathbf{r}} T}{T}}_{\text{Seebeck}} + \underbrace{K_{21}(\mathbf{B}\times\frac{\nabla_{\mathbf{r}} T}{T})}_{\text{Nernst}} + \underbrace{K_{31}\mathbf{B}(\mathbf{B}\cdot\frac{\nabla_{\mathbf{r}} T}{T})}_{\dots}],
\end{aligned} \tag{21.53}$$

where the important generalized kinetic coefficients K_{rs} have been defined[44]. The corresponding generalized heat current is

$$\begin{aligned}
\mathbf{J}_q &= \frac{1}{L^d}\sum_{\mathbf{k}}(E_{\mathbf{k}} - \mu)\mathbf{v}_{\mathbf{k}}\left(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}\right)(-\tau_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}\cdot\mathbf{G}) \\
\implies \mathbf{J}_q &= [\underbrace{K_{11}\mathbf{F}}_{\text{Peltier}} + \underbrace{K_{21}(\mathbf{B}\times\mathbf{F})}_{\text{Ettingshausen}} + \underbrace{K_{31}\mathbf{B}(\mathbf{B}\cdot\mathbf{F})}_{\dots}] \\
&+[\underbrace{K_{12}\frac{\nabla_{\mathbf{r}} T}{T}}_{-\kappa_e\nabla_{\mathbf{r}} T} + \underbrace{K_{22}(\mathbf{B}\times\frac{\nabla_{\mathbf{r}} T}{T})}_{\text{Righi-Leduc}} + \underbrace{K_{32}\mathbf{B}(\mathbf{B}\cdot\frac{\nabla_{\mathbf{r}} T}{T})}_{\dots}].
\end{aligned} \tag{21.54}$$

The kinetic coefficients[45] for *both* charge current $\mathbf{J}$ in Equation 21.53 and heat current $\mathbf{J}_q$ in 21.54 takes the unified form

$$\boxed{K^{ij}_{rs} = \frac{1}{L^d}\cdot\left(\frac{Q}{m^\star}\right)^{r-1}\cdot\sum_{\mathbf{k}}\left(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}\right)\frac{v_i v_j \tau^r_{\mathbf{k}}(E_{\mathbf{k}} - \mu)^s}{1 + (\frac{Q\tau_{\mathbf{k}}}{m^\star})^2|\mathbf{B}|^2},} \tag{21.55}$$

where the subindices (r, s) appear as powers of the relaxation time τ and energy $(E_{\mathbf{k}} - \mu)$ inside the sum and (i, j) capture its tensor nature.

---

[42] The total electric field $\mathbf{E} = \mathbf{E}_{\text{int.}} + \mathbf{E}_{\text{ext}} = \nabla_{\mathbf{r}}(\frac{\mu}{q} - \varphi)$ is the sum of the externally applied electric field $\mathbf{E}_{\text{ext}} = -\nabla_{\mathbf{r}}\varphi$ where φ is the scalar electric potential, and internal quasi-electric forces due to heterostructures or spatially nonuniform doping $\mathbf{E}_{\text{int.}} = -\nabla_{\mathbf{r}}\frac{\mu}{(-q)} = \nabla_{\mathbf{r}}(\frac{E_F - E_c}{q})$ that may be present.

[43] Equation 21.52 is worked out in Exercise 21.9. The general solution with an effective mass tensor represented by a matrix $[M]$ instead of the scalar $m^\star$ is provided there. Note that for charge Q, the force $\mathbf{F} = Q\mathbf{E}$ where $\mathbf{E}$ is the electric field.

[44] The physical meaning of the terms and their vector driving forces are explicitly seen here. For example, the charge current flows parallel to the electric field $\mathbf{E}$, but the Hall-current flows along $\mathbf{B}\times\mathbf{E}$, perpendicular to $\mathbf{B}$ and $\mathbf{E}$. A charge current also flows in response to a temperature gradient in the absence of $\mathbf{E}$ and $\mathbf{B}$: this is the Seebeck effect, and so on...

[45] With this generalized formula we can calculate experimentally measurable macroscopic kinetic coefficients from the microscopic bandstructure and scattering processes for transport in any dimension. The same formula is applicable to calculate the electrical and thermal conductivities, Hall effect and magnetoresistance, thermoelectric coefficients due to the Seebeck and Peltier effects, and several galvanomagnetic and thermomagnetic effects such as the Nernst-Ettingshausen and Righi-Leduc effects, thus encompassing the rich range of transport phenomena exhibited in semiconductors and their quantized heterostructures. This chapter covers transport when $\mathbf{B} = 0$, and Chapter 25 focuses on magnetotransport.

The L^d in the denominator indicates the volume in d-dimensions that cancels upon summation over $\mathbf{k}$. Inside the sum, $v_i = \frac{1}{\hbar}(\nabla_{\mathbf{k}} E_{\mathbf{k}}) \cdot \hat{\mathbf{k}}_i$ and $v_j = \frac{1}{\hbar}(\nabla_{\mathbf{k}} E_{\mathbf{k}}) \cdot \hat{\mathbf{k}}_j$ are the group velocity components in the i and j directions obtained from the bandstructure, $\hat{\mathbf{k}}_i$ is the unit vector in the i direction, $Q = -q$ is the electron charge and $m^\star$ is the isotropic effective mass[46]. The last three kinetic coefficients of the particle current K_{11}, K_{21}, and K_{31} are identical to the first three of the heat current, even though they *seem* to be different phenomena. These are the famous Onsager relations in action, as we discuss shortly.

21.8 Electrical conductivity

Let us consider the simplest situation in the absence of magnetic fields ($\mathbf{B} = 0$) and temperature gradients ($\nabla_r T = 0$), with an electric field $\mathbf{E} = \mathbf{E}_{\text{int.}} + \mathbf{E}_{\text{ext}} = \nabla_{\mathbf{r}}(\frac{\mu}{q} - \varphi)$ as the lone force[47]. Before using the coefficients K^{ij}_{rs} of Equation 21.55, a direct connection to experimentally measured electrical conductivities due to drift and diffusion processes is possible from the RTA solution of Equation 21.42 with a $\mathbf{k}$-dependent $\tau_{\mathbf{k}}$. For the electrical (charge) current $\mathbf{J}$ the physical property associated with each $\mathbf{k}$ state is $\phi_{\mathbf{k}} = -q\mathbf{v}_{\mathbf{k}}$, with $\mathbf{v}_{\mathbf{k}} = \hbar^{-1}\nabla_{\mathbf{k}} E_{\mathbf{k}}$ the group velocity of state $\mathbf{k}$ and $-q$ the electron charge identical for all $\mathbf{k}$. The current is

$$\mathbf{J} = \frac{1}{L^d}\sum_{\mathbf{k}}(-q\mathbf{v}_{\mathbf{k}})f_{\mathbf{k}} \approx -\frac{q}{L^d}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}(f^0_{\mathbf{k}} + \tau_{\mathbf{k}}\mathbf{v}_{\mathbf{k}} \cdot [(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})\mathbf{F} + (-\nabla_{\mathbf{r}} f^0_{\mathbf{k}})])$$

$$= \underbrace{-\frac{q}{L^d}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}} f^0_{\mathbf{k}}}_{=0} + \underbrace{\frac{q}{L^d}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}} \cdot \mathbf{F})\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}\tau_{\mathbf{k}}}_{\mathbf{J}_{\text{drift}}} + \underbrace{\frac{q}{L^d}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}} \cdot \nabla_{\mathbf{r}} f^0_{\mathbf{k}})\tau_{\mathbf{k}}}_{\mathbf{J}\text{ diffusion}}, \tag{21.56}$$

where three distinct components naturally result from the RTA:

1. **"Zero" current:** The first current term is zero by symmetry because the equilibrium Fermi–Dirac distribution function is even in $\mathbf{k}$ ($f^0_{\mathbf{k}} = f^0_{-\mathbf{k}}$), but the group velocity $\mathbf{v}_{\mathbf{k}}$ is odd ($\mathbf{v}_{\mathbf{k}} = -\mathbf{v}_{-\mathbf{k}}$), the sum of their product over all $\mathbf{k}$ is zero. However, we must not forget that the individual "right"-going component $\mathbf{J}^0_R$ and "left"-going component $\mathbf{J}^0_L$ of the current are not zero, but equal in magnitude to $J_0 = (-q)/L^d \sum_{\mathbf{k}>0} \mathbf{v}_{\mathbf{k}} f^0_{\mathbf{k}} \cdot \hat{\mathbf{n}}$ where $\hat{\mathbf{n}}$ is a unit vector in any direction[48].

2. **Drift current:** Consider the force $\mathbf{F} = -q\mathbf{E}$ on an electron due to an electric field $\mathbf{E}$ alone. With the group velocity $\mathbf{v}_{\mathbf{k}} = \frac{1}{\hbar}\nabla_{\mathbf{k}} E_{\mathbf{k}}$, the general drift current $\mathbf{J}_{\text{drift}}$ in Equation 21.56 reads $\mathbf{J}_{\text{drift}} = [\sigma]\mathbf{E}$, where

---

[46] Exercise 21.9 generalizes the kinetic coefficient to anisotropic effective masses represented by the tensor matrix $[M]$.

[47] Since charge current and heat current (and a temperature gradient) are intertwined by Equations 21.53 and 21.54, this cannot be *exactly* correct. What is meant here is that there is no *external* temperature gradient imposed on the system.

[48] This "internal" current J_0 may be evaluated for any bandstructure and in any direction. It is not due to external fields or forces, but due to thermal energy in the case of non-degenerate carriers, or due to the Pauli exclusion principle for degenerate carriers. Since these currents are equal and opposite in all directions, they cancel to give a net zero current. These directional currents were seen to play a major role in ballistic transport in several earlier chapters.

the electrical conductivity $[\sigma]$ is a tensor:

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = \frac{q^2}{\hbar^2 L^d} \sum_{\mathbf{k}} (-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}) \cdot \tau_{\mathbf{k}} \cdot \begin{bmatrix} (\frac{\partial E_{\mathbf{k}}}{\partial k_x})^2 & \frac{\partial E_{\mathbf{k}}}{\partial k_x}\frac{\partial E_{\mathbf{k}}}{\partial k_y} & \frac{\partial E_{\mathbf{k}}}{\partial k_x}\frac{\partial E_{\mathbf{k}}}{\partial k_z} \\ \frac{\partial E_{\mathbf{k}}}{\partial k_y}\frac{\partial E_{\mathbf{k}}}{\partial k_x} & (\frac{\partial E_{\mathbf{k}}}{\partial k_y})^2 & \frac{\partial E_{\mathbf{k}}}{\partial k_y}\frac{\partial E_{\mathbf{k}}}{\partial k_z} \\ \frac{\partial E_{\mathbf{k}}}{\partial k_z}\frac{\partial E_{\mathbf{k}}}{\partial k_x} & \frac{\partial E_{\mathbf{k}}}{\partial k_y}\frac{\partial E_{\mathbf{k}}}{\partial k_y} & (\frac{\partial E_{\mathbf{k}}}{\partial k_z})^2 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix},$$

$$\implies \text{for } \mathbf{E} = \begin{bmatrix} E_x \\ 0 \\ 0 \end{bmatrix}, \boxed{J_x = [\underbrace{\frac{q^2}{\hbar^2 L^d} \sum_{\mathbf{k}} (-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}) \cdot \tau_{\mathbf{k}} \cdot (\frac{\partial E_{\mathbf{k}}}{\partial k_x})^2}_{\sigma_{xx} = qn\mu_n}] \cdot E_x.} \tag{21.57}$$

The drift current density J_x in response to the electric field is thus simply the scalar conductivity σ_{xx} times the field E_x. This conductivity can be obtained for *any* bandstructure $E_{\mathbf{k}}$ if the scattering time $\tau_{\mathbf{k}}$ is known. The conductivity is further split[49] into the product $\sigma_{xx} = qn\mu$ of electron charge with the mobile carrier density n and a **drift mobility** μ_n. The microscopic scattering processes contribute to the conductivity via the $\tau_{\mathbf{k}}$ and the bandstructure through $E_{\mathbf{k}}$.

3. **Diffusion current:** From Equation 21.56 the diffusion current is driven by the real-space gradient of the equilibrium distribution function $\nabla_{\mathbf{r}} f^0_{\mathbf{k}}$. It is therefore written as $\mathbf{J}_{\text{diff}} = [D]\nabla_{\mathbf{r}} f^0_{\mathbf{k}}$ with a tensor diffusion constant $[D]$, which in terms of all components is:

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = \frac{q}{\hbar^2 L^d} \sum_{\mathbf{k}} \tau_{\mathbf{k}} \cdot \begin{bmatrix} (\frac{\partial E_{\mathbf{k}}}{\partial k_x})^2 & \frac{\partial E_{\mathbf{k}}}{\partial k_x}\frac{\partial E_{\mathbf{k}}}{\partial k_y} & \frac{\partial E_{\mathbf{k}}}{\partial k_x}\frac{\partial E_{\mathbf{k}}}{\partial k_z} \\ \frac{\partial E_{\mathbf{k}}}{\partial k_y}\frac{\partial E_{\mathbf{k}}}{\partial k_x} & (\frac{\partial E_{\mathbf{k}}}{\partial k_y})^2 & \frac{\partial E_{\mathbf{k}}}{\partial k_y}\frac{\partial E_{\mathbf{k}}}{\partial k_z} \\ \frac{\partial E_{\mathbf{k}}}{\partial k_z}\frac{\partial E_{\mathbf{k}}}{\partial k_x} & \frac{\partial E_{\mathbf{k}}}{\partial k_y}\frac{\partial E_{\mathbf{k}}}{\partial k_y} & (\frac{\partial E_{\mathbf{k}}}{\partial k_z})^2 \end{bmatrix} \begin{bmatrix} \frac{\partial f^0_{\mathbf{k}}}{\partial x} \\ \frac{\partial f^0_{\mathbf{k}}}{\partial y} \\ \frac{\partial f^0_{\mathbf{k}}}{\partial z} \end{bmatrix},$$

$$\implies \text{for } \nabla_{\mathbf{r}} f^0_{\mathbf{k}} = \begin{bmatrix} \frac{\partial f^0_{\mathbf{k}}}{\partial x} \\ 0 \\ 0 \end{bmatrix}, \boxed{J_x = q \cdot \sum_{\mathbf{k}} \tau_{\mathbf{k}} \cdot \underbrace{\frac{1}{\hbar^2} \cdot (\frac{\partial E_{\mathbf{k}}}{\partial k_x})^2}_{v^2\tau \sim D} \cdot \underbrace{(\frac{1}{L^d} \cdot \frac{\partial f^0_{\mathbf{k}}}{\partial x})}_{\sim \frac{dn}{dx}}.} \tag{21.58}$$

The diffusion current is obtained if $E_{\mathbf{k}}$ and $\tau_{\mathbf{k}}$ are known, both of which are identical to their values for the drift current. This microscopic connection manifests as the macroscopic Einstein relation, which is seen to emerge immediately if we connect the formulations of drift and diffusion currents in Equations 21.57 and 21.58 to the generalized coefficients in Equation 21.55. For $\mathbf{B} = \nabla_{\mathbf{r}} T = 0$, only the first term survives[50] in Equation 21.53, which in terms of coefficient K_{10} gives:

$$\mathbf{J} = -qK_{10}\mathbf{E} = -qK_{10}\nabla_{\mathbf{r}}(\frac{\mu}{q} - \varphi) = \underbrace{q^2 K_{10}\mathbf{E}_{\text{ext.}}}_{qn\mu_n\mathbf{E}_{\text{ext.}}} + \underbrace{qK_{10}(-\nabla_{\mathbf{r}}\mu)}_{qD_n\nabla_{\mathbf{r}} n}$$

$$\implies \boxed{\mu_n = \frac{qK_{10}}{n} \text{ and } D_n = K_{10}\frac{-\nabla_{\mathbf{r}}\mu}{\nabla_{\mathbf{r}} n} \implies \frac{D_n}{\mu_n} = \frac{k_b T}{q}.} \tag{21.59}$$

[49] This splitting is not just a matter of convenience, but of direct physical consequence in semiconductors. In diodes and transistors the mobile carrier density n changes by several orders of magnitude while in operation, whereas the mobility μ_n hardly changes at all.

[50] In the non-degenerate approximation $n = N_c \exp[(E_F - E_c)/k_b T] = N_c \exp[\mu/k_b T]$ which gives $\nabla_{\mathbf{r}}\mu = (k_b T/n)\nabla_{\mathbf{r}} n$ for the diffusion component. The case of general degeneracy is given in Equation 21.62. The Einstein relation $D/\mu = k_b T/q$ is an example of a broad class of *fluctuation-dissipation theorems* of statistical mechanics.

Mobilities, anisotropic bandstructures, and scattering: The above formulations indicate currents could also flow in directions in which there are no forces! But this only occurs for anisotropic energy bandstructures of particular kinds. For ellipsoidal bandstructures of type $E_{\mathbf{k}} = \hbar^2/2(k_x^2/m_{xx} + k_y^2/m_{yy} + k_x^2/m_{zz})$ which is the diagonal case of

$$E_{\mathbf{k}} = \frac{\hbar^2}{2}[\mathbf{k}]^T[M]^{-1}[\mathbf{k}] = \frac{\hbar^2}{2}[k_x\ k_y\ k_z]\begin{bmatrix} \frac{1}{m_{xx}} & \frac{1}{m_{xy}} & \frac{1}{m_{xz}} \\ \frac{1}{m_{yx}} & \frac{1}{m_{yy}} & \frac{1}{m_{yz}} \\ \frac{1}{m_{zx}} & \frac{1}{m_{zy}} & \frac{1}{m_{zz}} \end{bmatrix}\begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix}, \quad (21.60)$$

the off-diagonal terms of drift and diffusion currents in Equations 21.57 and 21.58 of the form $(\partial E_{\mathbf{k}}/\partial k_x)(\partial E_{\mathbf{k}}/\partial k_y) = \hbar^4 k_x k_y/(m_{xx}m_{yy})$ vanish upon summation over all $\mathbf{k}$ on account of being odd around the k_x, k_y axes. The electrical conductivity and mobility (and the related diffusion constant) and the kinetic coefficient K_{10}^{ij} are therefore diagonal (and simpler!) for ellipsoidal bandstructures[51]. For electrons of density $n = \frac{1}{L^d}\sum_{\mathbf{k}} f_{\mathbf{k}}^0$ free to move in d-dimensions, they are

$$\boxed{\sigma_d = q^2 K_{10} = \frac{q^2}{L^d}\sum_{\mathbf{k}}(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}})v^2\tau_{\mathbf{k}} \implies \mu_d = \frac{\sigma_d}{qn} = \frac{\frac{q^2}{L^d}\sum_{\mathbf{k}}(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}})v^2\tau_{\mathbf{k}}}{\frac{q}{L^d}\sum_{\mathbf{k}} f_{\mathbf{k}}^0}.} \quad (21.61)$$

Because the d-dimensional carrier concentration is $n = N_c F_{\frac{d}{2}-1}(\eta)$, the generalized Einstein relation for any degeneracy becomes[52]

$$\mu_n = \frac{qK_{10}}{N_c F_{\frac{d}{2}-1}(\eta)} \text{ and } D_n = K_{10}\frac{k_b T}{N_c F_{\frac{d}{2}-2}(\eta)} \implies \boxed{\frac{D_n}{\mu_n} = \frac{k_b T}{q}\frac{F_{\frac{d}{2}-1}(\eta)}{F_{\frac{d}{2}-2}(\eta)}.} \quad (21.62)$$

A non-degenerately doped semiconductor at 300 K of mobility $\mu_n = 1000$ cm^2/V·s has a diffusion constant $D_n = \mu_n \frac{k_b T}{q} = 26$ cm^2/s. For $n = 10^{17}$/cm^3 and $|\mathbf{E}_{\text{ext.}}| = 1$ kV/cm the drift current density is $J_{drift} = 16$ kA/cm^2. For a concentration gradient of $\Delta n = 10^{17}$/cm^3 over a distance $\Delta x = 10^{-4}$ cm, the diffusion current is $J_{diff.} = 4.16$ kA/cm^2.

To proceed further in the evaluation of electron mobilities, it is necessary to know the energy bandstructure $E_{\mathbf{k}}$ and the scattering time $\tau_{\mathbf{k}}$ of the specific microscopic processes. For parabolic bandstructure $E_{\mathbf{k}} = \hbar^2 k^2/2m^\star$, using the kinetic coefficient K_{10}, a generalized expression for electron mobility for transport in d-dimensions is obtained[53]:

$$\boxed{\mu_d = \frac{q}{m^\star}\frac{\int_0^\infty du \frac{\tau_u \cdot u^{\frac{d}{2}}}{\cosh^2[\frac{u-\eta}{2}]}}{2\Gamma(\frac{d}{2})F_{\frac{d}{2}-1}(\eta)} = \frac{q\tau_0}{m^\star}\frac{\Gamma(\frac{d}{2}+p+1)}{\Gamma(\frac{d}{2}+1)}\frac{F_{\frac{d}{2}+p-1}(\eta)}{F_{\frac{d}{2}-1}(\eta)}.} \quad (21.63)$$

For the case of $d = 3$ and $\eta \gg +1$, representing highly degenerate electron gases similar to a metal, the above expressions simplify on account of the Dirac delta function, giving $\sigma \sim \frac{nq^2\tau_F}{m^\star}$ and $\mu \sim \frac{q\tau_F}{m^\star}$, results

[51] This is indeed the case for most conduction bands: Si and Ge have ellipsoidal conduction band minima. Most III-Vs have spherical minima. Such luxury is not available for valence bands: for them spherical/ellipsoidal minima are a poor approximation.

[52] In the non-degenerate approximation $\eta \ll -1$ the Fermi–Dirac integrals $F_j(\eta) \approx e^\eta$, which reduces to Equation 21.59. For the degenerate approximation $\eta \gg +1$, the Einstein relation is $\frac{D_n}{\mu_n} = \frac{2}{d}\cdot\frac{(E_F - E_c)}{q}$ for diffusion in d-dimensions.

[53] The identity $(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}}) = \frac{1}{4k_b T\cosh^2[\frac{E_{\mathbf{k}}-\mu}{2k_b T}]}$ is used here with $u = E_{\mathbf{k}}/k_b T$ and $\eta = \mu/k_b T = (E_F - E_c)/k_b T$. For highly degenerate cases with $\eta \gg +1$, $(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}}) \approx \delta(E_{\mathbf{k}} - \mu)$, and for non-degenerate cases with $\eta \ll -1$, $(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}}) \approx \frac{e^\eta}{k_b T}$. If, as shown in Chapters 22 and 23, the scattering time can be expressed as $\tau_{\mathbf{k}} = \tau_0(\frac{E_{\mathbf{k}}}{k_b T})^p$, the expression for the mobility is obtainable in terms of Gamma functions and the Fermi–Dirac integrals on the right of Equation 21.63.

similar to the Drude model for a metal. For non-degenerate situations and the general case, a knowledge of $\mathbf{k}$- or energy-dependence of $\tau_{\mathbf{k}}$ is necessary. This is as far as one can proceed without this knowledge. The evaluation of individual scattering rates $\tau_{\mathbf{k}}$ expressed as τ_u in Equation 21.63 is the topic of Chapter 22 for phonon scattering and Chapter 23 for defect and impurity scattering[54].

For the rest of this section, we discuss general features, and the nature of τ that is obtained from the RTA. We focus on three types of relaxation times: the quantum τ_q, momentum τ_m, and energy τ_E relaxation times that represent physically different time constants. Before providing a derivation, they are simple to state and to understand:

$$\begin{aligned} \frac{1}{\tau_q(\mathbf{k})} &= \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}') && \text{[Quantum Relaxation Rate]} \\ \frac{1}{\tau_m(\mathbf{k})} &= \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')\left(1 - \frac{\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}'}}{\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}}}\right) && \text{[Momentum Relaxation Rate]} \\ \frac{1}{\tau_E(\mathbf{k})} &= \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')\left(1 - \frac{E_{\mathbf{k}'}}{E_{\mathbf{k}}}\right) && \text{[Energy Relaxation Rate].} \end{aligned} \tag{21.64}$$

The relaxation rate for state $\mathbf{k}$ is a sum over the scattering "out" transition rates $S(\mathbf{k} \to \mathbf{k}')$ from that state to all other possible states $\mathbf{k}'$, obtained by Fermi's golden rule separately for each type of scattering mechanism. The quantum and energy relaxation rates are properties of the $\mathbf{k}$ states alone. The momentum scattering rate depends on $\mathbf{k}$, and can also depend on the direction of the electric field $\mathbf{E}$ in anisotropic bandstructures. The forms in Equation 21.64 can be suitably amended to account for the Pauli exclusion principle when needed, as will be demonstrated when we use them. In the presence of scattering, the quantum relaxation rate is never zero. If the momentum represented by the group velocity $\mathbf{v}_{\mathbf{k}}$ does not change upon scattering, then the momentum relaxation rate is zero for that event. Similarly, if the energy does not change in a scattering event, $E_{\mathbf{k}} = E_{\mathbf{k}'}$ makes the energy relaxation rate zero, which occurs for elastic scattering events.

Fig. 21.15 illustrates the scattering times. Due to "quantum" scattering, after time τ_q the electron distribution loses memory that they were all pointing in *exactly* the same direction, but still have memory of the *net* initial momentum. The distance propagated in this time is $l_q \sim v\tau_q$. After the momentum relaxation time τ_m and a distance $l_m \sim v\tau_m$, the memory of the net initial momentum is erased, but the electron system still has the memory of the initial energy as represented by the lengths of the arrows. After the energy relaxation time τ_E and length $l_E \sim v\tau_E$, the electrons have scattered and lost their initial energy, and come to equilibrium with the crystal and the defects and vibrations. Typically $\tau_q < \tau_m < \tau_E$, and they are typically of the order of picosecond at room temperature. These values can change significantly depending on the situation. Let us now directly obtain the momentum relaxation rate from the RTA solution of the BTE.

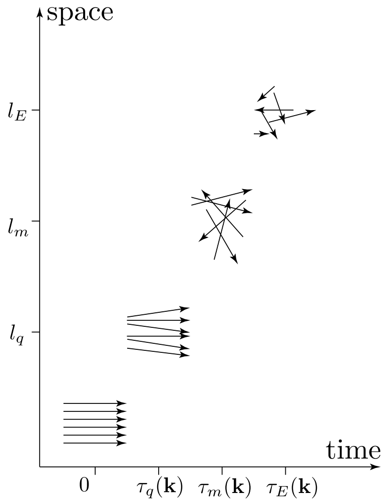


Fig. 21.15 The three relaxation times shown schematically in a space-time diagram for several electrons injected into a semiconductor as a collimated beam at time $t = 0$. The direction and length of the arrows represent the momentum vector for each electron. The quantum, momentum, and energy relaxation times, and the corresponding mean-free paths are indicated.

---

[54] For example in a 2D electron gas of $m^\star = 0.067$ and scattering time $\tau_F \sim 0.3$ ps, the mobility is $\sim 8,000$ cm^2/V·s and the sheet-resistance $1/\sigma$ is ~ 800 $\Omega/\Box$, fairly representative of 2DEGs in AlGaAs/GaAs heterostructures at room temperature, where phonon scattering limits the τ_F. At low temperatures τ_F can increase dramatically in clean samples from $\tau_F \sim 0.3$ ps to $\tau_F \sim 300$ ps or more as phonons are frozen out, boosting the mobility to $8,000,000$ cm^2/V·s, and reveal rich quantum transport phenomena such as quantized conductance and the quantum Hall effect, discussed in Chapter 25.

For **elastic scattering** the electron energy $E_{\mathbf{k}} = E_{\mathbf{k}'}$ is the same before and after scattering and therefore $S(\mathbf{k}' \to \mathbf{k}) = S(\mathbf{k} \to \mathbf{k}')$ from the principle of microscopic reversibility. The collision term in the RHS of the BTE $\frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}} = -\sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')[f_{\mathbf{k}} - f_{\mathbf{k}'}]$, rewritten as

$$\frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}} + f_{\mathbf{k}}\underbrace{[\sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')]}_{1/\tau_q(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}') f_{\mathbf{k}'}, \tag{21.65}$$

shows the meaning of the **quantum relaxation rate** as the sum of the scattering rates $1/\tau_q(\mathbf{k}) = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')$ from a state $\mathbf{k}$ to all other states $\mathbf{k}'$. The corresponding quantum scattering time $\tau_q(\mathbf{k})$ may be viewed as a "lifetime" of the particle in the state $|\mathbf{k}\rangle$. A particle in this state $|\mathbf{k}\rangle$ at time $t = 0$ will scatter out into other states $|\mathbf{k}'\rangle$ due to collisions. The occupation function $f_{\mathbf{k}}$ of that state will approach the equilibrium value $f^0_{\mathbf{k}}$ exponentially fast as $\sim \exp[-t/\tau_{\mathbf{q}}(\mathbf{k})]$ with the time constant $\tau_q(\mathbf{k})$ when external fields or perturbations are removed.

Consider the steady state case in the presence of an electric field that produces a force $-q\mathbf{E}$, making the solution of the BTE in the RTA $f_{\mathbf{k}} = f^0_{\mathbf{k}} + q\tau_{\mathbf{k}} \mathbf{v}_{\mathbf{k}} \cdot \mathbf{E} \frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}$. For elastic scattering $|\mathbf{k}| = |\mathbf{k}'|$ which makes $f^0_{\mathbf{k}} = f^0_{\mathbf{k}'}$ and $\tau_{\mathbf{k}} = \tau_{\mathbf{k}'}$. The momentum relaxation rate is then obtained:

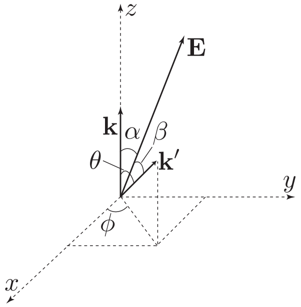


Fig. 21.16 Angular relations between $\mathbf{k}$, $\mathbf{k}'$, and $\mathbf{E}$ vectors in the Boltzmann transport equation.

$$\frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}} = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')[f_{\mathbf{k}'} - f_{\mathbf{k}}] \;\&\; f_{\mathbf{k}} - f_{\mathbf{k}'} = \underbrace{q\tau_{\mathbf{k}} \mathbf{v}_{\mathbf{k}} \cdot \mathbf{E} \frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}}_{f_{\mathbf{k}} - f^0_{\mathbf{k}}} [1 - \frac{\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}'}}{\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}}}]$$

$$\implies \frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}} = -\frac{(f_{\mathbf{k}} - f^0_{\mathbf{k}})}{\tau_m(\mathbf{k})} \text{ where } \frac{1}{\tau_m(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')[1 - \frac{\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}'}}{\mathbf{E} \cdot \mathbf{v}_{\mathbf{k}}}]. \tag{21.66}$$

Since $f_{\mathbf{k}} - f^0_{\mathbf{k}}$ does not depend on $\mathbf{k}'$, it is possible to pull it out of the summation, leading to a clean definition of the momentum relaxation rate in the elastic scattering approximation. The formulation further simplifies if the angular relation between the three vectors: the group velocities $\mathbf{v}_{\mathbf{k}}$, $\mathbf{v}_{\mathbf{k}'}$ and the electric field $\mathbf{E}$ are considered, as shown in Fig. 21.16 for a parabolic bandstructure, when $\mathbf{v}_{\mathbf{k}} = \hbar\mathbf{k}/m^\star$ and $\mathbf{v}_{\mathbf{k}'} = \hbar\mathbf{k}'/m^\star$. Fixing the z-axis along $\mathbf{k}$ and the y-axis so that $\mathbf{E}$ lies in the $y - z$ plane, from Fig. 21.16 the relation

$$\frac{\mathbf{k}' \cdot \mathbf{E}}{\mathbf{k} \cdot \mathbf{E}} = \frac{\cos\beta}{\cos\alpha} = \cos\theta + \sin\theta \sin\phi \tan\alpha \tag{21.67}$$

is obtained (this is the addition theorem of spherical harmonics). When the sum over all final states $\mathbf{k}'$ is performed, the term containing $\sin\phi$ vanishes by symmetry, simplifying the momentum relaxation time $\tau_m(\mathbf{k})$ to a form that is used heavily in evaluating scattering rates:

$$\frac{1}{\tau_m(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}')(1 - \cos\theta), \tag{21.68}$$

where θ is the angle between $\mathbf{k}$ and $\mathbf{k}'$. Though derived in 3D, it holds for other dimensions. The momentum relaxation time is what determines the electrical conductivity and mobility via $\mu = q\langle\tau_m\rangle/m^\star$ suitably averaged over all states $\langle\tau_m\rangle$ as derived in Equation 21.63. The $1-\cos\theta$ weights lead to a large value for back-scattering when $\theta \to \pi$, making τ_m small and thus lowering the mobility. For small angle scattering when $\theta \to 0$, the factor can be much smaller than unity: this increases τ_m and results in a higher mobility[55].

The quantum scattering rate $1/\tau_q(\mathbf{k}) = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')$ and the momentum scattering rate $1/\tau_m(\mathbf{k}) = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')(1-\cos\theta)$ are both experimentally accessible quantities[56], and provide a valuable method to identify the nature of scattering mechanisms. The momentum scattering time $\tau_m(\mathbf{k})$ measures the average time spent by the particle moving along the external field. It differs from the quantum lifetime due to the $\cos\theta$ term. For scattering processes that are isotropic, $S(\mathbf{k},\mathbf{k}')$ has no angle dependence, the $\cos\theta$ term sums to zero when integrated over all θ, and $\tau_q = \tau_m$. However, for scattering processes that favor small angle ($\theta \to 0$) scattering, from the definitions, $\tau_m > \tau_q$. For example, Coulomb potentials heavily favor small angle scattering, whereas short-range scatterers such as uncharged point defects are isotropic scatterers, as discussed further in Chapter 23. The difference between the quantum and momentum scattering times is used to distinguish between different scattering mechanisms in experiments.

21.9 Thermoelectric properties

Fig. 21.17 schematically explains how a temperature gradient across a semiconductor drives an electrical current, or generates a voltage across it. The energy band diagram shows that the Fermi function f_h has a longer high-energy tail in the hot region on the left side, and a smaller tail in f_c for the cold region on the right. There is a preferential motion of electrons at higher energies to the right, each transporting an energy of amount $E_{\mathbf{k}} - \mu \sim k_b T_h$ from the left to the right side, and a charge of magnitude q. Similarly, an energy of amount $k_b T_c$ is transported from the right to the left. If the electrons were to accumulate in the electrodes at the two ends, a voltage difference of magnitude $V \sim \frac{k_b}{q}(T_h - T_c) = \frac{k_b}{q}\Delta T$ builds up, which is $\sim 86\ \mu$V for $\Delta T = 1$ K. The sign of the voltage is sensitive to the direction of the temperature gradient and the charge of the carrier: it can distinguish between n- and p-type semiconductors. If the left and right sides are brought closer in space, the same voltage difference will lead to a larger internal electric field proportional to the temperature gradient $\frac{\Delta T}{\Delta x}$. Completing the outer circuit will drive a current proportional to the temperature gradient, $J \sim \frac{dT}{dx}$. The Fermi function difference $f_h - f_c$ in Fig. 21.17 (c) shows that most of the thermoelectric current is carried by electrons close to $E_{\mathbf{k}} \geq \mu = E_F - E_c$. Designing large DOS in this energy window is a strategy to maximize thermoelectric effects.

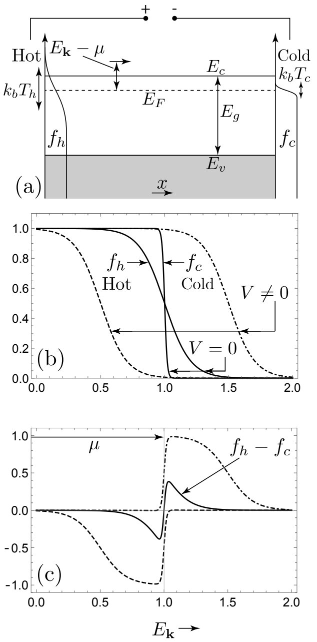


Fig. 21.17 (a): The energy band diagram, (b): the distribution functions f_h, f_c, and (c): the driving force for thermoelectric transport, the difference of the distribution functions $f_h - f_c$. The semiconductor with band edges E_c, E_v, and energy gap E_g is hot on the left side and cold on the right, creating a temperature gradient $\frac{dT}{dx} = \nabla_{\mathbf{r}} T$. The solid lines of (b) and (c) are when there is a temperature gradient but no externally applied voltage difference. The dashed and dot-dashed lines indicate f_h, f_c and $f_h - f_c$ when an external voltage difference is also present in addition to the temperature gradient.

---

[55] Examples are discussed in detail in Chapters 22 and 23 for various scattering mechanisms.

[56] Typically the momentum scattering time is concluded by either a Hall-effect or field-effect measurement of the mobility. The quantum scattering time is obtained in low-temperature magnetotransport quantum oscillations such as the Shubnikov–de-Haas oscillations of longitudinal magnetoresistance, as discussed in Chapter 25.

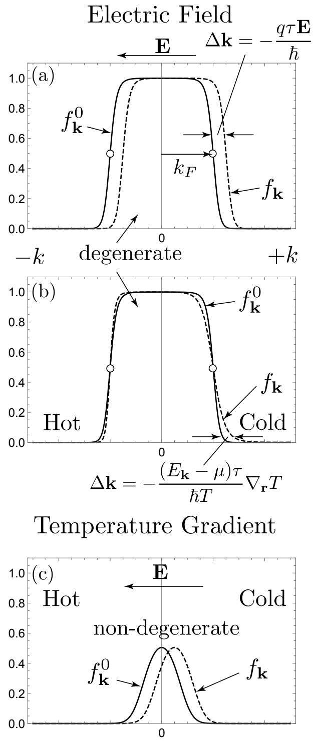


Fig. 21.18 Difference between the effects of an electric field (a) and thermal gradient (b) on the distribution function for degenerate electron distributions. In (c) is the case for non-degenerate distribution. In this case because the Fermi level is in the energy gap, $E_{\mathbf{k}} - \mu > 0$ which does not switch sign. Note that $\mu = E_F - E_c$ for the conduction band. The labels of "Hot" and "Cold" are in real space, so the carriers with $+k$ wavevector are from the left (Hot) side.

These qualitative thermoelectric features are now investigated quantitatively using the RTA solution of the BTE. Equations 21.53 and 21.54 show that the three forces that drive charge and heat current are the electric field $\mathbf{E}$, temperature gradients $\nabla_{\mathbf{r}}T$, and magnetic field $\mathbf{B}$. In the last section, we investigated only the charge current flow in response to an electric field assuming $\mathbf{B} = \nabla_{\mathbf{r}}T = 0$. For thermoelectric effects we retain the temperature gradient $\nabla_{\mathbf{r}}T$ but keep $\mathbf{B} = 0$ in this chapter (Chapter 25 covers thermoelectric phenomena for $\mathbf{B} \neq 0$).

The electro-thermal force vector for $\mathbf{B} = 0$ from Equation 21.52 is $\mathbf{G} = -q\mathbf{E} + (\frac{E_{\mathbf{k}}-\mu}{T})\nabla_{\mathbf{r}}T$. Consider first the effects of electric field and thermal gradient separately in the RTA solution $f_{\mathbf{k}} = f^0_{\mathbf{k}} + \tau_{\mathbf{k}}\mathbf{v}_{\mathbf{k}} \cdot \mathbf{G}(\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})$. Fig. 21.18 (a) shows the RTA solution of the BTE for a degenerate equilibrium electron distribution $f^0_{\mathbf{k}}$, which shifts by $\Delta\mathbf{k} = -q\tau\mathbf{E}/\hbar$ in the $\mathbf{k}$-space in response to an electric field $\mathbf{E}$. Fig. 21.18 (b) shows that the response to a temperature gradient $\nabla_{\mathbf{r}}T$ is different. The $\Delta\mathbf{k}$ is obtained by expressing the RTA solution as a Taylor expansion of the type $f_{\mathbf{k}} = f^0_{\mathbf{k}} - \Delta\mathbf{k} \cdot \nabla_{\mathbf{k}} f^0_{\mathbf{k}}$ to get

$$f_{\mathbf{k}} = f^0_{\mathbf{k}} + \frac{\tau(E_{\mathbf{k}}-\mu)}{T}\mathbf{v} \cdot \nabla_{\mathbf{r}}T\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}} \implies \Delta\mathbf{k} = -\frac{(E_{\mathbf{k}}-\mu)\tau}{\hbar T}\nabla_{\mathbf{r}}T, \tag{21.69}$$

which for small $\Delta\mathbf{k}$ follows $f_{\mathbf{k}} = f^0_{\mathbf{k}} - \Delta\mathbf{k} \cdot \nabla_{\mathbf{k}} f^0_{\mathbf{k}} \approx f^0_{\mathbf{k}}(\mathbf{k} - \Delta\mathbf{k})$. This means the equilibrium Fermi–Dirac distribution is shifted by $\Delta\mathbf{k}$. But for a temperature gradient, $\Delta\mathbf{k}$ from Equation 21.69 is dependent on the *sign* of $E_{\mathbf{k}} - \mu = E_{\mathbf{k}} - (E_F - E_c)$ for the conduction band for example. The case for a degenerate electron distribution is shown in Fig. 21.18 (b). The temperature gradient $\nabla_{\mathbf{r}}T < 0$ because the left side is hot. Then, for right-going electrons $k > 0$ and energies above the Fermi level $E_{\mathbf{k}} \geq \mu$, the sign is $\Delta\mathbf{k} > 0$. For right-going electrons and $E_{\mathbf{k}} \leq \mu$, $\Delta\mathbf{k} < 0$, and all the other shifts of $f_{\mathbf{k}}$ from the equilibrium value $f^0_{\mathbf{k}}$ can be reasoned in this way. By thus breaking the symmetry of $f^0_{\mathbf{k}}$, the temperature gradient leads to a net flow of electrons from the left to the right, and a charge current flow from the right to the left. If the semiconductor is non-degenerate, the electron energies $E_c + E_{\mathbf{k}} > E_F$ implies $E_{\mathbf{k}} - \mu > 0$ at all $\mathbf{k}$ values, and thus the distribution function shifts in a manner similar to the electric field, as shown in Fig. 21.18 (c). This is a crucial difference between degenerate and non-degenerate semiconductors. Loosely speaking, thermoelectric effects are stronger for non-degenerate semiconductors.

Since charge and heat currents are coupled, the thermoelectric effects are obtained directly from the coupled charge and heat transport Equations 21.53 and 21.54, which for $\mathbf{B} = 0$ reduce to

$$\mathbf{j} = \frac{\mathbf{J}}{(-q)} = K_{10}\mathbf{E} + K_{11}\frac{\nabla_{\mathbf{r}}T}{T}, \text{ and } \mathbf{J}_q = K_{11}\mathbf{E} + K_{12}\frac{\nabla_{\mathbf{r}}T}{T}. \tag{21.70}$$

In experiments, it is easier to control and measure an electric current $\mathbf{J}$ than the internal electric field $\mathbf{E}$. Expressing $\mathbf{E}$ and the heat current $\mathbf{J}_q$

in terms of the current **J** and the temperature gradient gives

$$\mathbf{E} = \underbrace{(\frac{1}{-qK_{10}})}_{\rho}\mathbf{J} + \underbrace{(-\frac{K_{11}}{K_{10}T})}_{S}\nabla_{\mathbf{r}}T \implies \boxed{\mathbf{E} = \rho\mathbf{J} + S\nabla_{\mathbf{r}}T,} \text{ and}$$

$$\mathbf{J}_q = \underbrace{(\frac{K_{11}}{-qK_{10}})}_{\Pi}\mathbf{J} + \underbrace{(\frac{K_{12} - \frac{K_{11}^2}{K_{10}}}{T})}_{-\kappa_e}\nabla_{\mathbf{r}}T \implies \boxed{\mathbf{J}_q = \Pi\mathbf{J} - \kappa_e\nabla_{\mathbf{r}}T,} \quad (21.71)$$

where the boxed equations define the four transport coefficients: the **electrical resistivity** $\rho = \frac{1}{\sigma}$, the **Seebeck coefficient**[57] S, the **Peltier coefficient** Π, and the **electronic thermal conductivity** κ_e. The electrical resistivity is the inverse of the electrical conductivity which was discussed in the last section. We focus on the interrelatedness of these three thermoelectric coefficients, and their practical applications.

Seebeck coefficient: Writing the charge current as $\mathbf{J} = \sigma(-\nabla_{\mathbf{r}}V) - \sigma S\nabla_{\mathbf{r}}T$, and from Fig. 21.17 (a) considering the open-circuit situation when $\mathbf{J} = 0$, the Seebeck coefficient becomes $S = -\frac{\Delta V}{\Delta T}$, the voltage difference resulting from a temperature difference. This is how it is experimentally measured. The values reached in lightly doped semiconductors around room temperature can be as high as $S \sim 10\frac{k_b}{q} \sim 1$ mV/K, but in heavily doped semiconductors and metals it is an order of magnitude lower. The reason becomes clear if the Seebeck coefficient is directly calculated using $S = -K_{11}/K_{10}T$ from Equation 21.71 and the generalized kinetic coefficients from Equation 21.55:

$$S = -\frac{K_{11}}{qK_{10}T} = -\frac{1}{qT}\frac{\sum_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})v^2_{\mathbf{k}}\tau_{\mathbf{k}}(E_{\mathbf{k}} - \mu)}{\sum_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})v^2_{\mathbf{k}}\tau_{\mathbf{k}}} = -\frac{\langle E_{\mathbf{k}} - \mu\rangle}{qT}, \quad (21.72)$$

where the angular bracket $\langle E_{\mathbf{k}} - \mu\rangle$ indicates the kind of weighted averaging indicated by the sum over all **k** states. This is a generalization of what is commonly referred to as the Mott formula for the Seebeck coefficient. Evaluating the above sum for parabolic bands in d-dimensions assuming $\tau_{\mathbf{k}} = \tau_0(\frac{E_{\mathbf{k}}}{k_bT})^r$, we get for the non-degenerate $\eta \ll -1$ and degenerate $\eta \gg +1$ limits (see Exercise 21.11):

$$\boxed{S \underbrace{\approx}_{\eta\ll-1} -\frac{k_b}{q}[\underbrace{\frac{E_c - E_F}{k_bT}}_{-\eta} + \frac{d+2}{2} + r]} \text{ and } \boxed{S \underbrace{\approx}_{\eta\gg+1} -\frac{k_b}{q}[\frac{d}{2} + r]\frac{\pi^2}{3}\underbrace{\frac{k_bT}{E_c - E_F}}_{-1/\eta}}. \quad (21.73)$$

Fig. 21.19 (a) shows the Seebeck coefficient and (b) shows the electrical conductivity as a function of the Fermi level. When the Fermi level is deep inside the bandgap, the Seebeck coefficient is high[58], but the electrical conductivity is low because there are very few carriers. The electrical conductivity increases when the Fermi level enters the conduction band due to a large increase in carrier density, but the Seebeck

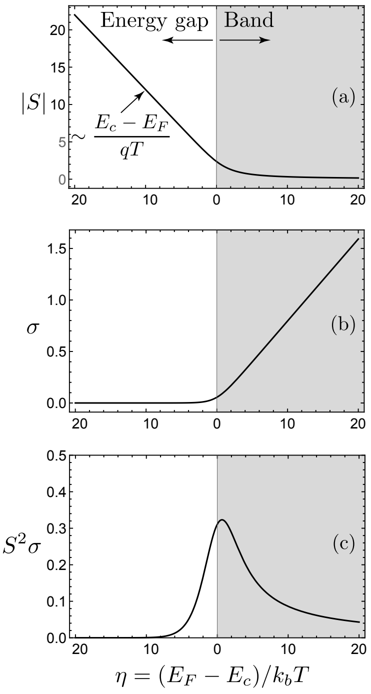


Fig. 21.19 (a) The Seebeck coefficient S, (b) the electrical conductivity σ, and the power factor $S^2\sigma$ for a semiconductor plotted in dimensionless units as a function of $\eta = (E_F - E_c)/k_bT$, the location of the Fermi level with respect to the conduction band edge.

[57] To avoid excessive symbols we use S for the Seebeck coefficient assuming that the context of usage distinguishes it from the scattering rate $S(\mathbf{k} \to \mathbf{k}')$ and entropy $\mathcal{S}$. Though it must be mentioned that the Seebeck coefficient has a deeper connection to entropy: in magnitude it is the entropy per unit charge $S = \mathcal{S}/nq$ where n is the carrier density. The Seebeck coefficient is also called the **thermopower**.

[58] For non-degenerate semiconductors, the Seebeck coefficient is dominated by the $\eta = (E_F - E_c)/k_bT$ term, which can be much larger than the $\frac{d+2}{2} + r$ term which is of the order of unity. For this case one can approximate $|S| \approx \frac{k_b}{q}|\eta|$. The general expression for all values of η is $S = -\frac{k_b}{q}[-\eta + (\frac{d+2}{2} + r)\frac{F_{r+\frac{d}{2}}(\eta)}{F_{r+\frac{d}{2}-1}(\eta)}]$.

coefficient plummets because k_bT is small compared to the degeneracy $E_F - E_c$. The product $S^2\sigma$ thus has a maximum when the Fermi level is located close to the band-edge. This product turns up as the power factor in thermoelectric materials, as discussed shortly.

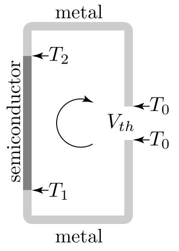


Fig. 21.20 Seebeck and thermocouple effect. The voltage V_{th} that develops at the electrode terminals is proportional to the temperature difference $T_2 - T_1$.

The Seebeck effect is the principle of operation of **thermocouples** that produce a voltage proportional to a junction temperature. The thermocouple arrangement also allows the measurement of the Seebeck coefficient of semiconductors. Consider Fig. 21.20, in which a semiconductor is connected to two metal contacts in an open circuit. Both metal leads are made of the same metal, and the terminals are held at the same temperature T_0. If the two metal/semiconductor contacts are at temperatures T_1 and T_2, a thermal voltage of

$$\begin{aligned}-V_{th} = \oint S dT &= \int_{T_0}^{T_1} S_{\text{met}} dT + \int_{T_1}^{T_2} S_{\text{semi}} dT + \int_{T_2}^{T_0} S_{\text{met}} dT \\ &= \int_{T_1}^{T_2} [S_{\text{semi}} - S_{\text{met}}] dT \approx S_{\text{semi}}(T_2 - T_1) \quad (21.74)\end{aligned}$$

develops across the metal terminals, which can be measured and tabulated for various $T_2 - T_1$ to give the Seebeck coefficient of the semiconductor S_{semi}. The integral is taken in the clockwise loop as indicated in Fig. 21.20. Typically the Seebeck coefficient of semiconductors is much larger than that of metals S_{met}, but the approximation $S_{\text{semi}} - S_{\text{met}} \approx S_{\text{semi}}$ is not necessary if the Seebeck coefficient of the metal is known.

Peltier coefficient: Closely related to the Seebeck effect is the fact that the flow of an electrical current, in addition to transporting charge, also transports heat, even in the *absence* of an externally enforced temperature gradient. This is the Peltier effect, which relates the heat current to the charge current via Equation 21.71, which for $\nabla_{\mathbf{r}} T = 0$ is $\mathbf{J}_q = \Pi \mathbf{J}$, where the **Peltier coefficient** $\Pi = \frac{K_{11}}{K_{10}} = ST$ is simply the Seebeck coefficient times the temperature with the unit of volts[59].

Fig. 21.21 shows a charge current $\mathbf{J}$ flowing across a junction of materials of different Peltier coefficients. Because $\mathbf{J}_{q1} = \Pi_1 \mathbf{J}$ and $\mathbf{J}_{q2} = \Pi_2 \mathbf{J}$, we have $\mathbf{J}_{q1} - \mathbf{J}_{q2} = \Delta \mathbf{J}_q = (\Pi_1 - \Pi_2)\mathbf{J}$. This implies that a Peltier coefficient difference will produce heating at the junction when the charge current flows in one direction. But when the charge current flows in the opposite direction, because $\Delta \mathbf{J}_q$ changes sign, the junction will experience cooling! This is the principle of Peltier heating and cooling. Equations 21.50 and 21.71 give the net generated heat as

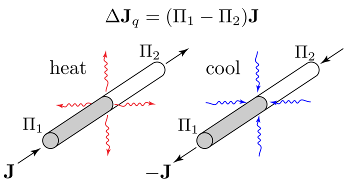


Fig. 21.21 Peltier heating and cooling.

$$Q = \mathbf{J} \cdot \mathbf{E} - \nabla_{\mathbf{r}} \cdot \mathbf{J}_q = \mathbf{J} \cdot [\frac{1}{\sigma}\mathbf{J} + S \nabla_{\mathbf{r}} T] - \nabla_{\mathbf{r}} \cdot [\Pi \mathbf{J} - \kappa_e \nabla_{\mathbf{r}} T]$$

$$= \underbrace{\frac{|\mathbf{J}|^2}{\sigma}}_{\text{Joule heat}\downarrow} - \underbrace{T\frac{dS}{dT}\mathbf{J} \cdot \nabla_{\mathbf{r}} T}_{\text{Thomson heat}\uparrow} + \underbrace{\nabla_{\mathbf{r}} \cdot [\kappa_e (\nabla_{\mathbf{r}} T)]}_{\text{Thermal cond.}\downarrow}, \quad \boxed{ZT = \frac{S^2 \sigma T}{\kappa_e + \kappa_l},} \quad (21.75)$$

where the net heat produced is split into three terms. The first term is the familiar **Joule heat**, and the third term is the heat due to the

[59] That the Peltier and Seebeck coefficients are related in this way is one of the Onsager relations which is discussed in the next section. For non-degenerate semiconductors, the Peltier coefficient is simply $|\mathbf{J}_q|/|\mathbf{J}| = \Pi = ST \approx \frac{E_F - E_c}{q}$, which quantitatively verifies the prior qualitative discussion that $E_c - E_F$ of heat is transported by charge $-q$.

electronic part of the thermal conductivity gradient. The second term called the **Thomson heat** is unconventional: unlike Joule heat, which is always positive, the Thomson heat switches sign when the current switches direction. This heat is either released, or absorbed from the environment depending on the direction of the current! This is rather remarkable, and can be put to clever use for practical purposes, such as a single device that operates as a **Peltier refrigerator *and* oven**, as shown in Fig. 21.22. The principle of operation is similar to Fig. 21.21, but now there are several junctions, and n- and p-type semiconductors are used to boost the Peltier coefficient difference. The most efficient operation will be achieved if the Thomson term in Equation 21.75 is maximized, and the Joule and thermal conductivity terms are minimized. This can be achieved with a high Seebeck coefficent, a large electrical conductivity, and a low thermal conductivity, and are embodied in the factor ZT of Equation 21.75. The power factor $S^2\sigma$ appears in the numerator, and the *total* thermal conductivity in the denominator that includes the electronic (κ_e) and lattice (κ_l) parts.

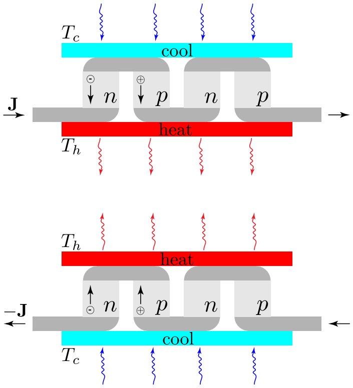


Fig. 21.22 A semiconductor based solid-state Peltier cooler and heater. Current flowing in one direction heats the upper surface and cools the lower, and vice versa upon switching the current direction. The metal/semiconductor/metal structures are connected in series, the semiconductors alternating between n- and p-type. The top and bottom metal electrodes are in thermal contact with different plates. The device is electrically in series, but thermally in parallel.

That ZT is an important thermoelectric figure of merit is seen by writing a simplified version of Equation 21.75 for one of the legs of the Peltier cooler in Fig. 21.22 of length L, cross-section A, thermal conductivity κ and resistance $R = \frac{1}{\sigma}\frac{L}{A}$ as $Q_c LA = P_c = I^2R - SIT + \kappa\frac{A}{L}\Delta T$, where we have used $\frac{dS}{dT} \sim \frac{S}{T}$ for a non-degenerate semiconductor. The heat extraction from the cold side is maximized for the current I for which $\partial P_c/\partial I = 0$, which is $I = ST/2R$. Substituting back we get $P_c^{\max} = \kappa\frac{A}{L}\Delta T - \frac{S^2T^2}{4R}$, which when zero maximizes the temperature difference between the hot and cold sides. This difference $\Delta T = \frac{1}{4}\cdot\frac{S^2\sigma}{\kappa}\cdot T^2$ indicates that maximizing the factor $Z = \frac{S^2\sigma}{\kappa}$ indeed maximizes the thermoelectric performance of the device.

If instead of flowing a current, the device was placed between a hot and cold surface, a voltage will be generated across the two leads, converting the temperature difference $T_h - T_c$ into a voltage, making it a **thermoelectric battery**. The use of the alternating n- and p-semiconductors will then, via the Seebeck effect, generate a voltage $V_n = -S_n(T_h - T_c)$ and $V_p = -S_p(T_c - T_h)$ across each stage, and for $N\times$ n-p stages, they will add in series to produce an open circuit voltage of magnitude $|V_{oc}| \sim N(T_h - T_c)(|S_n| + |S_p|)$.

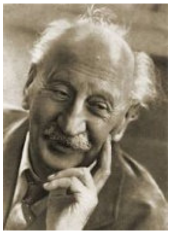


Fig. 21.23 Abram Ioffe, the physicist who early on recognized the potential of semiconductors for thermoelectric power conversion, among several other contributions to physics.

***ZT*, electronic and thermal conductivity, and Lorenz number:** To investigate the variation of ZT in various materials, we discuss the electronic and thermal conductivities. In the presence of a temperature gradient, if there is no charge current flow ($\mathbf{J} = 0$), Equation 21.71 shows that there is still a heat current $\mathbf{J}_q = -\kappa_e\nabla_{\mathbf{r}}T$ carried by electrons. In metals, most of the heat current is carried by electrons, but in semiconductors, a large portion of the heat current is carried by lattice vibrations or phonons (discussed in Chapter 22), and electrons contribute only a very small part due to their low density. This small portion defines the **electronic part** of the thermal conductivity which in terms of the generalized kinetic coefficients is $\kappa_e = (K_{11} - K_{11}^2/K_{10})/T$.

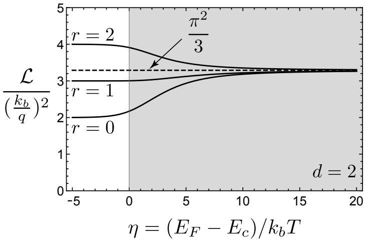


Fig. 21.24 Lorenz number for electron transport in $d = 2$ dimensions.

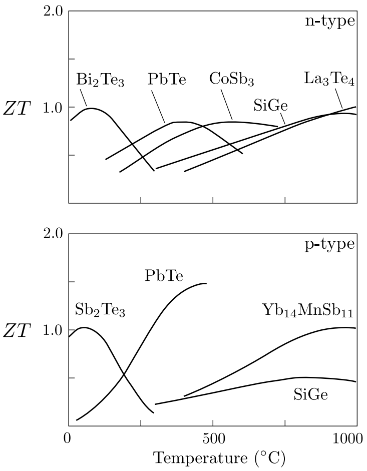


Fig. 21.25 Thermoelectric ZT values vs. temperature in °C for various n-type and p-type thermoelectrics.

Just as for the classical Drude model in Chapter 2, the ratio of the electronic part of the thermal conductivity and the electronic conductivity defines a **Lorenz number** (see Exercise 21.12 for the exact form):

$$\mathcal{L} = \frac{\kappa_e}{\sigma T} = \frac{1}{q^2T^2}\left[\frac{K_{12}}{K_{10}} - \left(\frac{K_{11}}{K_{10}}\right)^2\right] = \frac{\langle (E_{\mathbf{k}} - \mu)^2\rangle - \langle E_{\mathbf{k}} - \mu\rangle^2}{q^2T^2} \implies$$

$$\boxed{\mathcal{L} \underbrace{\approx}_{\eta \ll -1} \left(r + \frac{d+2}{2}\right) \cdot \left(\frac{k_b}{q}\right)^2} \quad \text{and} \quad \boxed{\mathcal{L} \underbrace{\approx}_{\eta \gg +1} \frac{\pi^2}{3} \cdot \left(\frac{k_b}{q}\right)^2,} \qquad (21.76)$$

where d is the dimensions the charge moves in, and r as before for the Seebeck coefficient appears as the energy exponent of the scattering time $\tau_{\mathbf{k}} = \tau_0(\frac{E_{\mathbf{k}}}{k_bT})^r$. Fig. 21.24 shows the Lorenz number for electron transport in two-dimensions which occurs in MOSFETs and in 2D electron gases in semiconductor quantum wells. When the carrier density is degenerate, the Lorenz number becomes $\frac{\pi^2}{3}(\frac{k_b}{q})^2$ which is independent of the scattering mechanisms and also the dimension. This follows the Wiedemann–Franz law which is similar, but not quite in the same way as the classical formulation as discussed in Chapter 2. On the other hand for the non-degenerate case the scattering mechanism and the dimensionality both determine the asymptotic value of $(r + \frac{d+2}{2})(\frac{k_b}{q})^2$ approached for $\eta \ll -1$.

A large ZT is thus desirable for both Peltier cooling/heating as well as a thermoelectric battery. Using the Lorenz number $\mathcal{L}$, the ZT of any material is $ZT = \frac{S^2}{\mathcal{L}}$, which is dimensionless. The value of ZT of metals is low, because they have low S and high κ_e which outweighs their high σ. Since the thermal conductivity in metals is nearly entirely due to electrons, $ZT \sim \frac{S^2}{\mathcal{L}} \sim \frac{\pi^2}{3}(\frac{k_bT}{\mu})^2 \ll 1$, making them unsuitable for thermoelectric energy conversion, or cooling/heating. Semiconductors and semimetals on the other hand boast higher ZT values, some close to unity. The search for materials that have $ZT \sim 2-3$ continues, and will have revolutionary impact if this is successful. The need of low κ and high σ makes heavy-atom semiconductors/semimetals from the lower part of the periodic table the most attractive (Fig. 21.25).

The currently available ZT values restrict the efficiencies of thermoelectric energy conversion efficiency to less than 10%. Because this is lower than heat engines based on moving gases or fluids, they have not made mainstream impact yet. Nevertheless, semiconductor and semimetal-based thermoelectrics have already found several uses that effectively exploit their special traits. Because of movement of electrons and holes (instead of molecules), they are compact, fast, and because they have no "moving parts", they have almost infinite lifetime. Some of the popular thermoelectric materials are doped Bi_2Te_3/Sb_2Te_3 and Si/SiGe superlattices. They are used for precision temperature control with fast feedback for controlling semiconductor lasers and other devices which are sensitive to temperature. They have enabled compact electric beverage heaters and coolers, car-seat and windscreen heaters, and have several automotive and space applications[60].

[60] A fascinating application is the **radioactive thermoelectric generator**, which sips off the nuclear fuel of radioactive materials released as heat, and converts it to electricity. They last as long as the half-life of the radioactive materials, and thus can be "self-powered" for 100s of years. They are thus used in places that are prohibitive for human life. Such thermoelectric generators have helped explore the parts of the solar system where solar energy is not available, and is one of the few human-built devices that has gone beyond the solar system in the Voyager spacecraft.

21.10 Onsager relations

In the preceeding sections we have seen that a generalized current in the direction i in response to a generalized force F_j in the direction j is a sum $J_i = \sum_j L_{ij} F_j$, which written in the matrix form is

$$\begin{bmatrix} J_1 \\ J_2 \\ J_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} L_{11} & L_{12} & L_{13} & \cdots \\ L_{21} & L_{22} & L_{23} & \cdots \\ L_{31} & L_{32} & L_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ \vdots \end{bmatrix}. \tag{21.77}$$

In Equation 21.51 we have seen that the corresponding rate of change of entropy is $\frac{dS}{dt} = \sum_i J_i F_i = \sum_{ij} F_i L_{ij} F_j$. Since the entropy should increase with time in the approach towards equilibrium $\frac{dS}{dt} > 0$, the square matrix of the kinetic coefficients L_{ij} should be **positive definite**, which means all its eigenvalues are positive. Onsager (Fig. 21.26) showed that when there is no magnetic field ($\mathbf{B} = 0$), in addition to being positive definite, the matrix of transport coefficients is also **symmetric**, that is $L_{ij} = L_{ji}$. In the presence of a magnetic field, time reversal symmetry implies that the relation $L_{ij}(+\mathbf{B}) = L_{ji}(-\mathbf{B})$ must hold. This is the Onsager relation, which he derived from statistical mechanics, and by invoking the reversibility of the microscopic laws of motion governing the particles of the system.

Fig. 21.26 Lars Onsager formulated the reciprocal relations for kinetic coefficients from fundamental thermodynamic arguments. He also provided the exact solution of the 2D Ising model and realized that quantum oscillations of magnetic susceptibility in the de-Haas–van Alphen effect as a means to measure the Fermi surfaces of metals. Onsager was awarded the 1968 Nobel Prize in Chemistry.

We encountered the first Onsager relation in thermoelectric phenomena. The Seebeck coefficient S and the Peltier coefficient Π were found to be connected by the relation $\Pi = TS$ via the kinetic coefficients[61] derived in Equations 21.53, 21.53, and 21.55, and then applied in Equation 21.71. This was already realized in Equations 21.53, where the kinetic coefficients of the last three charge current components related to the forces due to temperature gradients and magnetic fields were found to be exactly the same as the first three components of the heat current in Equation 21.54 due to electric and magnetic fields and not due to thermal forces[62]. Long before Onsager's derivation, the thermoelectric relation between the Seebeck and Peltier coefficients was known as the **Kelvin relation** after Lord Kelvin (William Thomson), who had derived and explained it from thermodynamic considerations.

The above matrix for charge current $\mathbf{J}$ and heat current $\mathbf{J}_q$ in three dimensions in response to the three forces $\mathbf{E}$, $\mathbf{B}$, and $\nabla_{\mathbf{r}} T$ thus forms a 6×6 L_{ij} matrix, characterized by 36 kinetic coefficients of the type in Equation 21.55. Because of Onsager's relations, the 15 off-diagonal matrix elements are symmetric, implying 21 independent kinetic coefficients. Because all currents are interrelated through this matrix, there are several more Onsager relations in the presence of magnetic fields, which go by the names of researchers that discovered them experimentally before Onsager explained why they were the same. The several Onsager relations in the presence of a magnetic field are discussed for thermo-magnetic transport phenomena in Chapter 25.

[61] If instead of defining the driving force as $\nabla_{\mathbf{r}} T$, we had used $\nabla_{\mathbf{r}}(\frac{1}{T})$, then the identity $L_{12} = L_{21}$ would be obtained.

[62] Thus, one can loosely say that the Onsager relations imply that the ratio of a electrical particle current to the temperature gradient that drives it, is identical to the ratio of the corresponding heat current to the electrical gradient that drives it. It connects phenomena that naively may seem distinct, but at the microscopic level are connected because of the reversibility of the dynamical laws that govern their motion.

21.11 Conservation laws

We have investigated the steady-state solution of the BTE in response to DC electric or magnetic fields, or temperature gradients. This is insufficient for forces that oscillate in time. A solution method for oscillating forces for the BTE will now be developed. In addition to its use in later chapters on magnetotransport and photonic properties of semiconductors, it will illuminate and strenghten features of the BTE by connecting to conservation laws that were encountered earlier.

The general principle is called taking "moments" of the BTE. Using the RTA solution for the collision term $\frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}} \approx -\frac{f_{\mathbf{k}}-f^0_{\mathbf{k}}}{\tau_0}$, multiplying the BTE by $\phi_{\mathbf{k}}$, a function of $\mathbf{k}$, and summing over all $\mathbf{k}$ gives

$$\frac{1}{\Omega}\sum_{\mathbf{k}}\phi_{\mathbf{k}}(\frac{\partial f_{\mathbf{k}}}{\partial t}+\mathbf{v}_{\mathbf{k}}\cdot\nabla_{\mathbf{r}}f_{\mathbf{k}}+\frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}}f_{\mathbf{k}}) = -\frac{1}{\Omega}\sum_{\mathbf{k}}\phi_{\mathbf{k}}(\frac{f_{\mathbf{k}}-f^0_{\mathbf{k}}}{\tau_0}) \implies$$

$$\boxed{\frac{\partial\rho_\phi(\mathbf{r},t)}{\partial t}+\frac{1}{\Omega}\sum_{\mathbf{k}}\phi_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}}\cdot\nabla_{\mathbf{r}}f_{\mathbf{k}}+\frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}}f_{\mathbf{k}}) = -\frac{1}{\Omega}\sum_{\mathbf{k}}\phi_{\mathbf{k}}(\frac{f_{\mathbf{k}}-f^0_{\mathbf{k}}}{\tau_0}),} \tag{21.78}$$

where the density $\rho_\phi(\mathbf{r},t) = \frac{1}{\Omega}\sum_{\mathbf{k}}\phi_{\mathbf{k}}f_{\mathbf{k}}(\mathbf{r},t)$ per volume Ω is the spatial variation of the physical property $\phi_{\mathbf{k}}$ at time t. For simplicity we assume a $\mathbf{k}$-independent τ_0 (more general solutions can be sought). Choosing $\phi_{\mathbf{k}}$ successively as different physical quantities generates various conservation laws, which are now described (See Table 21.1).

Table 21.1 Moments of the BTE and the corresponding conservation laws.

$\phi_{\mathbf{k}}$	Conservation Law
Particle#	
1	$\frac{\partial n}{\partial t}+\nabla_{\mathbf{r}}\cdot\mathbf{j}=0$
Energy	
$E_{\mathbf{k}}$	$\frac{\partial u_e}{\partial t}+\nabla_{\mathbf{r}}\cdot\mathbf{J}_{u_e}=\mathbf{F}\cdot\mathbf{J}-R$
Entropy	
$s_{\mathbf{k}}$	$\frac{\partial s_e}{\partial t}+\nabla_{\mathbf{r}}\cdot\mathbf{J}_{s_e}=G_{s_e}$

Conservation of particles: For $\phi_{\mathbf{k}} = 1$, $\rho_\phi(\mathbf{r},t) = \frac{1}{\Omega}\sum_{\mathbf{k}} f_{\mathbf{k}}(\mathbf{r},t) = n(\mathbf{r},t)$ is the particle density. Because the number of particles is unchanged $\sum_{\mathbf{k}} f_{\mathbf{k}} = \sum_{\mathbf{k}} f^0_{\mathbf{k}}$, the RHS of Equation 21.78 vanishes. The equation then becomes the particle conservation law $\frac{\partial n}{\partial t}+\nabla_{\mathbf{r}}\cdot\mathbf{j} = 0$, where $\mathbf{j} = \frac{1}{\Omega}\sum\mathbf{v}_k f_{\mathbf{k}}$ is the particle current density, and the equation is of current continuity[63]. The electron charge current is $\mathbf{J} = -q\mathbf{j}$.

Conservation of electron energy: Choosing $\phi_{\mathbf{k}} = E_{\mathbf{k}}$ as the electron energy eigenvalue from the known bandstructure, in the scattering term on the RHS $R = \sum_{\mathbf{k}} E_{\mathbf{k}}(\frac{f_{\mathbf{k}}-f^0_{\mathbf{k}}}{\tau_0})$ is the net rate of energy loss by the electron population to the lattice due to scattering. Denoting the electron energy density by $u_e = \frac{1}{\Omega}\sum_{\mathbf{k}} f_{\mathbf{k}}E_{\mathbf{k}}$, the energy current density by $\mathbf{J}_{u_e} = \frac{1}{\Omega}\sum_{\mathbf{k}} E_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}f_{\mathbf{k}}$, Equation 21.78 becomes[64]

$$\boxed{\frac{\partial u_e}{\partial t}+\nabla_{\mathbf{r}}\cdot\mathbf{J}_{u_e} = \mathbf{E}\cdot\mathbf{J}-R,} \text{ and } \boxed{\frac{\partial u_l}{\partial t}+\nabla_{\mathbf{r}}\cdot\mathbf{J}_{u_l} = R,} \tag{21.79}$$

which are the energy conservation relations for the electron system and the lattice system. $\mathbf{E}\cdot\mathbf{J}$ is the generation term, since it is the energy absorbed by the electron system from the electric field, and R is the energy recombination or loss term for electrons: it is delivered to

[63] Here we use the physical argument that $\sum_{\mathbf{k}}\nabla_{\mathbf{k}}\cdot(\frac{d\mathbf{k}}{dt}f_{\mathbf{k}}) = \sum_{\mathbf{k}}\frac{d\mathbf{k}}{dt}\cdot(\nabla_{\mathbf{k}}f_{\mathbf{k}}) + \sum_{\mathbf{k}} f_{\mathbf{k}}(\nabla_{\mathbf{k}}\cdot\frac{d\mathbf{k}}{dt}) = \sum_{\mathbf{k}}\frac{d\mathbf{k}}{dt}\cdot(\nabla_{\mathbf{k}}f_{\mathbf{k}}) = 0$, since there is no net outflow or "divergence" of particles in the $\mathbf{k}$-space from the Brillouin zone in intraband transport.

[64] Since there is no net outflow of energy in the $\mathbf{k}$-space from the Brillouin zone, we have $\sum_{\mathbf{k}}\nabla_{\mathbf{k}}\cdot(E_{\mathbf{k}}\frac{d\mathbf{k}}{dt}f_{\mathbf{k}}) = 0$, which gives $\sum_{\mathbf{k}} E_{\mathbf{k}}\frac{d\mathbf{k}}{dt}\cdot(\nabla_{\mathbf{k}}f_{\mathbf{k}}) = \mathbf{E}\cdot\mathbf{J}$ where $\mathbf{E}$ is the electric fields. This relation is used to derive the law of energy conservation for the electron system.

the lattice by collisions, producing another conservation relation for the lattice energy density u_l.

Conservation of entropy and total energy: The relation for entropy generation was already discussed in Equation 21.50, where several of the current derivations from the moments of the BTE were also encountered from a thermodynamic viewpoint. We can now get the bigger picture from the BTE as to how the total energy in the transport process is conserved. The two Equations in 21.79 account for the energy stored in the electron system u_e and the lattice u_l. To complete the story we need to account also for the energy density u_{EM} in the electromagnetic field that drives the current. The conservation law is directly obtained from Maxwell's equations by writing

$$\frac{\partial}{\partial t}[\underbrace{\mathbf{D}\cdot\mathbf{E}+\mathbf{H}\cdot\mathbf{B}}_{u_{EM}}]+\nabla_{\mathbf{r}}\cdot[\underbrace{\mathbf{E}\times\mathbf{H}}_{\mathbf{J}_{EM}}]=-\mathbf{E}\cdot\mathbf{J}, \tag{21.80}$$

where $\mathbf{D}$ is the displacement vector and $\mathbf{H}$ the magnetic induction, which define the standard energy per unit volume u_{EM} in the electromagnetic field. The energy current density of the EM field $\mathbf{J}_{EM}$ is the **Poynting vector**. The RHS is the energy lost by the EM field to the electrons in the lattice. Adding the LHS and RHS of Equations 21.79 and 21.80 gives the total conservation relation

$$\boxed{\frac{\partial}{\partial t}(\underbrace{u_e}_{\text{electron}}+\underbrace{u_L}_{\text{lattice}}+\underbrace{u_{EM}}_{\text{EM Field}})+\nabla_{\mathbf{r}}\cdot(\mathbf{J}_{u_e}+\mathbf{J}_{u_L}+\mathbf{J}_{u_{EM}})=0,} \tag{21.81}$$

which explicitly shows the division of energy into the electron, lattice, and the EM systems, and ensures that in the transport process the total energy is always conserved.

Conservation of momentum: Let us now derive a rather important and widely applicable result using the moment of $\phi_{\mathbf{k}}=\mathbf{v}_{\mathbf{k}}$, the group velocity of state $\mathbf{k}$ in the electron bandstructure. Multiplying the BTE of $f_{\mathbf{k}}$ with the corresponding $\mathbf{v}_{\mathbf{k}}$ with the same $\mathbf{k}$, dividing by the volume, using $(\frac{\partial f_{\mathbf{k}}}{\partial t})_{\text{coll.}}=-\frac{f_{\mathbf{k}}-f_0}{\tau_0}$, and summing over all $\mathbf{k}$, we get

$$\begin{aligned}&\frac{1}{\Omega}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}\Big(\frac{\partial f_{\mathbf{k}}}{\partial t}+\mathbf{v}_{\mathbf{k}}\cdot\nabla_{\mathbf{r}}\cdot f_{\mathbf{k}}+\frac{\mathbf{F}}{\hbar}\cdot\nabla_{\mathbf{k}}f_{\mathbf{k}}\Big)=-\frac{1}{\Omega}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}\Big(\frac{f_{\mathbf{k}}-f_0}{\tau_0}\Big)\\ &\implies\boxed{\frac{\partial\mathbf{J}}{\partial t}+\frac{\mathbf{J}}{\tau_0}=\frac{q}{\hbar\Omega}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}(\mathbf{F}\cdot\nabla_{\mathbf{k}}f_{\mathbf{k}})+\frac{q}{\Omega}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}}\cdot\nabla_{\mathbf{r}}f_{\mathbf{k}}),}\end{aligned} \tag{21.82}$$

where $\mathbf{J}=\frac{-q}{\Omega}\sum_{\mathbf{k}}f_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}$ is the electron current density and $\mathbf{F}=-q(\mathbf{E}+\mathbf{v}_{\mathbf{k}}\times\mathbf{B})$ is the Lorentz force. Steps described in Exercise 21.15 transforms this moment equation of the BTE to a rather important form for transport in response to forces oscillating in time:

$$\boxed{\frac{\partial \mathbf{J}}{\partial t} + \frac{\mathbf{J}}{\tau_0} = \frac{nq^2}{m^\star}\mathbf{E} - \frac{q}{m^\star}(\mathbf{J} \times \mathbf{B}) + \frac{qD}{\tau_0}\nabla_{\mathbf{r}} n,} \tag{21.83}$$

where the macroscopic charge current density **J** is related to the applied electric field **E**, the magnetic field **B**, and any existing gradients in the carrier density $\nabla_{\mathbf{r}} n$. This result is valid for a parabolic bandstructure of effective mass $m^\star$ and density n, and D is the diffusion constant. This generalized dynamical form of the drift-diffusion equation helps explain a host of electronic, optical, and magnetic phenomena in semiconductors, and is especially useful for sinusoidal forces.

We briefly discuss this time-dependent formulation of the BTE. Several features of transport phenomena in semiconductors may be immediately recognized from Equation 21.83. Consider the steady state reached with a DC electric field **E**, at no magnetic field and no concentration gradients. Then, the **drift current** $\mathbf{J} = \frac{nq^2\tau_0}{m^\star}\mathbf{E} = \sigma\mathbf{E}$, where $\sigma = qn\mu$ is the conductivity, and $\mu = \frac{q\tau}{m^\star}$ is the mobility. If there were no electric or magnetic fields, $\mathbf{J} = qD\nabla_{\mathbf{r}} n$ is the **diffusion current**. In the presence of a DC magnetic field **B** with zero electric field and no carrier density gradients, an oscillating current may be identified by writing $(i\omega + \frac{1}{\tau_0})J = -\frac{qB}{m^\star}J$, identifying the cyclotron frequency $\omega_c = \frac{qB}{m^\star}$. If $\mathbf{B} = 0$ and the electric field is oscillating in time, of the AC form $\mathbf{E} = \mathbf{E}_0 e^{i\omega t}$ such as due to an electromagnetic wave, then $(i\omega + 1/\tau_0)\mathbf{J}_0 e^{i\omega t} = \frac{nq^2}{m^\star}\mathbf{E}_0 e^{i\omega t}$ gives the AC current $\mathbf{J}(\omega) = \frac{(nq^2\tau_0/m^\star)}{1+i\omega\tau_0}\mathbf{E}_0$. This is identical to the Drude form that was used back in Exercise 2.5 to explain the reflectivity of metals[65].

21.12 Berry curvature correction

Before discussing the limitations of the applicability of the BTE as formulated and discussed in this chapter in the next section, we point out a fundamental change that needs to be introduced in crystals that have non-zero Berry curvatures[66] $\mathbf{\Omega}_{\mathbf{k}}$ in their bandstructure. From Equation 21.1 the velocity of an electron wavepacket is $\mathbf{v}(\mathbf{k}) = \frac{1}{\hbar}\nabla_k E_n(k) - \frac{\mathbf{F}}{\hbar} \times \mathbf{\Omega}_{\mathbf{k}}$. It turns out that the phase–space volume $\Delta V = \Delta r \Delta k$ in transport is *not* conserved for non-zero $\mathbf{\Omega}_{\mathbf{k}}$, violating the Liouville theorem[67]. The phase space volume changes to $\frac{\Delta V}{\Delta V(0)} = (1 + \frac{q}{\hbar}\mathbf{B}\cdot\mathbf{\Omega})$. It may seem that the whole framework of BTE discussed in this chapter may not apply for such materials. This matter is resolved by simply re-defining the density of states in the following prescription:

$$\sum_{\mathbf{k}}(...) \to \sum_{\mathbf{k}}(1 + \frac{q}{\hbar}\mathbf{B}\cdot\mathbf{\Omega}_{\mathbf{k}}) \times (...), \tag{21.84}$$

which rescales the phase space so that the results of the BTE remain valid as long as the extra factor that includes the Berry curvature and magnetic field are included in all summations to evaluate transport coefficients.

---

[65] We continue this discussion and use the dynamic solution of the BTE in Chapters 25 and beyond.

[66] The boldface $\mathbf{\Omega}_{\mathbf{k}}$ is the **k**-dependent Berry curvature vector, not to be confused by the scalar symbol Ω sometimes used for a volume.

[67] Such situations occur for graphene, topological insulators, magnetic semiconductors, and also when spin-orbit interaction is considered in traditional semiconductors. The phase–space volume evolves in time as $\frac{1}{\Delta V}\frac{d\Delta V}{dt} = -\frac{d}{dt}\ln[1 + \frac{q}{\hbar}\mathbf{B}\cdot\mathbf{\Omega}]$ for non-zero Berry curvature materials when there is a non-zero magnetic field.

21.13 Limitations of the BTE

The BTE for electrons in a band is derived from the Liouville equation which requires that the electron number be conserved in the band. Thus, physical processes that move electrons into, or out of bands are not captured in the standard BTE. These situations, which are encountered when electron populations in bulk semiconductors are subjected to strong excitations that pull their distribution in the phase space very far from equilibrium, are difficult to describe by the BTE without significant extensions. These extensions require interband generation recombination currents, and coupled BTEs for each band involved in the process. Table 21.2 show the DC electric and magnetic fields, and AC electric fields (optical excitation) that can cause interband transitions in semiconductors of bandgap E_g. The strong perturbation conditions are reached easily in narrow bandgap semiconductors. In materials that have band overlaps, such as semimetals, or those that have band degeneracies such as Dirac cones in graphene, special care must be taken and the BTE must be extended accordingly to account for the special situation that low-field inelastic scattering (or for semimetals, even elastic scattering!) can move electrons between different bands.

Table 21.2 Physical conditions outside the regime of standard Boltzmann Transport Equation solutions. Here E_g is the bandgap, E_F the Fermi level, and a the lattice constant. These cases are discussed in later chapters on tunneling, magnetotransport, and optical properties.

Condition	Limit
Zener tunneling:	
DC field	$qFa > \frac{E_g^2}{E_F}$
Magnetic tunneling:	
DC Magnetic Field	$\hbar(\frac{qB}{m^\star}) > \frac{E_g^2}{E_F}$
Optical excitation:	
AC field ($E_0 e^{i\omega t}$)	$\hbar\omega > E_g$

Semiconductor devices for which the BTE approach either fails (or is unnecessary) to explain the simplest experimental phenomena serve to illustrate its limitations. The first example where the BTE is unnecessary is (surprisingly) the earliest semiconductor device: the Schottky diode. As was described in Chapter 18, Section 18.1, the Schottky diode is essentially a ballistic device. Therefore, for a Schottky diode a transport model that neglects scattering captures the device behavior better than those that account for scattering. As opposed to Bethe's ballistic model described in Section 18.1, models developed by Mott and by Schottky accounted for scattering and used a drift-diffusion approach for electrons moving from the semiconductor to the metal. These models also reproduce the $J = J_0(e^{qV/k_bT} - 1)$ behavior, and are useful in high-voltage Schottky diodes with long depletion region thicknesses much larger than the mean free path of electrons $x_d >> \lambda_{mfp}$. For Schottky diodes that have $x_d \leq \lambda_{mfp}$, the splitting of the quasi-Fermi levels at the semiconductor surface is $E_{Fs} - E_{Fm} = qV$, which is much closer to the ballistic model than those that need scattering and a BTE treatment.

As opposed to the Schottky diode where the BTE can be used (but is unnecessary), there are semiconductor devices where it fails to explain the basic operation. An example of such a device is the resonant tunneling diode (RTD) discussed in Chapter 24. The reason the BTE is inapplicable in that situation is because from the get-go, the BTE solves for the probability density $f(x, k, t)$ of the electrons in the phase space.

The probability density, analogous to the intensity of an electromagnetic wave, is proportional to the absolute square of the wavefunction amplitude $f \propto |\psi|^2$. Because the phase differences between the electron wavefunctions are lost in the BTE, it cannot capture the physics that is born out of interference effects caused by phase differences. In a RTD, a negative differential resistance (NDR) is observed when there are two potential barriers in the path of an electron. Electron wave interference allows resonant states to transmit through the barrier with essentially unity transmission probability. Because this transmission is due to constructive interference of the phase of the electron wavefunction, the BTE is unable to capture its physics[68].

Transport processes in which band electrons are captured by deep level defects and then re-emitted is not modeled well by the BTE. A recombination/generation model is better suited for such purposes. A model due to Mott, called the variable range hopping is more popularly used for situations where several electrons are trapped in deep defects and hop from one defect state to another. Exercise 21.14 shows how this model accounts for transport phenomena in the presence of severe disorder, which occurs in some semiconductors of technological relevance.

If the external driving forces that affect transport also strongly modify the underlying density of states of the semiconductor, then the BTE is no more applicable. The first such case occurs for high electric fields that result in interband tunneling of electrons in semiconductors. The finite spatial overlap of the valence band and conduction band wavefunctions then leads to an oscillatory joint density of states[69]. The interband tunneling current flow therefore is outside the scope of ordinary BTE because the number of carriers in a band are not conserved, violating the requirement of conserved particle number in the Liouville theorem. Possible extension of the BTE for coupled multiband transport can partially address this problem. Chapter 24 is devoted to the discussion of tunneling transport in semiconductors.

The second case where the BTE fails is for magnetotransport phenomena at very high magnetic fields, which is the subject of Chapter 25. A magnetic field changes the density of states of a solid by the formation of characteristic bunching into Landau levels[70]. The separation of the Landau levels is $\hbar\omega_c$, where $\omega_c = \frac{qB}{m^\star}$ is the cyclotron frequency. At small magnetic fields when $\hbar\omega_c << k_b T$ or $\hbar\omega_c << \frac{\hbar}{\tau}$ where τ is a scattering time, the oscillations in the density of states of the semiconductor is "washed out" by thermal broadening of the carrier distribution, or by disorder induced broadening. In these scenarios, the BTE is applicable and therefore used to explain low-field magnetotransport phenomena in semiconductors such as the classical Hall effect, cyclotron resonance, and Faraday rotation. But when $\hbar\omega_c >> k_b T$ or $\hbar\omega_c >> \frac{\hbar}{\tau}$, quantum oscillations such as the Shubnikov–de-Haas oscillations and the quantum Hall effect are readily observed in semiconductors. For such transport phenomena, the BTE is inadequate, and a density matrix approach based on the Kubo formalism is necessary.

---

[68] Quantum mechanical extensions of the BTE which solve for the wavefunction amplitude, such as the Wigner formalism can explain the behavior of the RTD. The non-equilibrium Green's function (NEGF) approach based on the Landauer picture of transport starts from a fully ballistic model, to which adding scattering and dephasing which becomes increasingly difficult numerically. Capturing all details of the RTD physics at room temperature, remains a stringent testing ground for the validity and power of theories and models for quantum transport.

[69] The most vivid experimental manifestation of this is the Franz–Keldysh effect which is a photon-assisted interband tunneling process, discussed in Chapter 27, Section 27.11.

[70] This introduces quasi-periodic oscillations in the density of states. The periodicity in DOS appears because of the requirement that the orbital periodicity of electron wavefunctions also satisfy an additional periodicity introduced by cyclotron orbits due to the magnetic field.

We discuss this fascinating topic in Chapter 25.

The BTE is therefore most powerful in **near-equilibrium transport phenomena that do not contain effects of electron phase interference**. It is the preferred method to explain and compute low-field transport coefficients such as mobilities and diffusion coefficients. Most importantly, its power lies in accounting for scattering, and for small perturbations of electric field, magnetic field, heat, or light with equal ease. Furthermore, the BTE is the best method to show that deep down, these perturbations are interrelated, via the Onsager relations.

21.14 Chapter summary section

This is one of the longest chapters of the book! We went deeper into irreversible thermodynamics in this chapter, and learned:

- How the Boltzmann transport equation (BTE) is obtained in the single-particle approximation of the Liouville equation.
- That by letting go of the knowledge of every electron and treating the transport problem statistically, the BTE makes it possible to discuss the effect of scattering of large number of electrons.
- The BTE shows that the equilibrium distribution of electrons is the one that has the highest entropy.
- The collision term of the BTE accounts for all microscopic scattering events via Fermi's golden rule.
- The relaxation-time approximation (RTA) is a linear solution to the BTE that allows for analytical solutions.
- Charge currents $\mathbf{J}$ and heat currents $\mathbf{J}_q$ flow in response to electric fields $\mathbf{E}$, magnetic fields $\mathbf{B}$, and temperature gradients $\nabla_{\mathbf{r}} T$.
- The kinetic coefficients that relate the currents to the forces are not independent: they are related by the Onsager relations.

Further reading

Richard Tolman's masterpiece **The Principles of Statistical Mechanics** provides a rigorous treatment of the Boltzmann H-theorem and irreversibility in classical and quantum mechanics. **The Boltzmann Equation and Its Applications** by Carlo Cercignani gives a comprehensive discussion of the Boltzmann equation and its many uses. In **Transport Phenomena**, Smith and Jensen shine the spotlight on the power and versatility of the Boltzmann transport equation by showing its use in understanding transport phenomena of gases, metals, semiconductors, insulators, and superconductors in a unified manner.

Various transport coefficients and the role of entropy is discussed in illuminating ways in **Electronic Conduction in Solids** by Smith, Janak, and Adler. Wolfe, Holonyak, and Stillman's **Physical Properties of Semiconductors**, and Lundstrom's **Fundamentals of Carrier Transport** provide in-depth discussion of the Boltzmann transport equation specifically for semiconductors.

Exercises

(21.1) **Why Bloch oscillations are difficult**

(a) Show that the time required for a Bloch oscillation is $t_B = \frac{h}{qFa} = \frac{100\text{ ps}}{(a \text{ in } 0.4\text{nm})\cdot(F \text{ in kV/cm })}$.

(b) To make the time comparable to the sub-ps scattering processes, one can increase the electric field. Estimate the field at which the time for Bloch oscillations is reduced to 0.1 ps.

(c) What problems would one run into at these high fields that would prevent Bloch oscillations for the single electron?

(21.2) **The Boltzmann trick**

To prove his rather famous H-theorem, Boltzmann used an ingenious physical argument to show that the expression for the rate of change of entropy of a spontaneous process must be mathematically always positive, implying that entropy increases, which is his proof of the second law of thermodynamics. In the step going from Equation 21.19 to Equation 21.20, show that exchanging $\mathbf{k} \rightarrow \mathbf{k_1}$ and the other indices sucessively should leave all of the physics of the electron-electron scattering unaltered, since they are dummy variables. Thereby, prove how Equation Equation 21.20 is obtained from Equation 21.19.

(21.3) **The arrow of time**

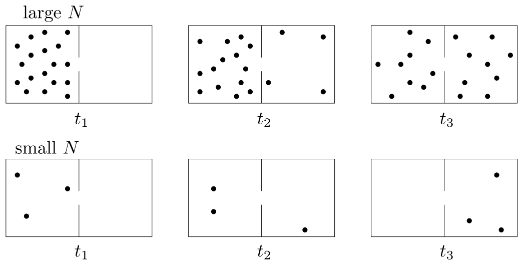


Fig. 21.27 A box with partition illustrates the relation of time's arrow and Boltzmann's H-theorem with fluctuations and statistical nature rooted in large numbers.

It is sometimes said that the second law of thermodynamics and direction of change of entropy defines an "arrow" of time. Fig. 21.27 shows three snapshots in time at $t = t_1, t_2$, and t_3 of a partitioned box in which there are gas molecules (or it could be an electron gas system in a solid with a split-gate partition).

(a) For the upper row for large N, what is the time-ordering of t_1, t_2, t_3? If your answer is $t_1 < t_2 < t_3$, then you have implicitly accepted the second law of thermodynamics. The states of the system have defined the direction of time, or time's arrow.

(b) What is the ordering of time for the lower row with small N?

(c) Discuss and comment on the statistical nature of the second law and time's arrow. At what N is there a crossover?

(d) Note that for the first row, in addition to large N, the state on the left is highly improbable, meaning it is a far from equilibrium low entropy state. Since the system cannot spontaneously end up there, someone must have put in the work to create this state. Argue then that time's arrow and Boltzmann's H-theorem also requires a low entropy initial state in addition to the large N.

(e) Argue why as semiconductor devices move increasingly into the nanoscale, a reconsideration of several cherished laws of thermodynamics (that require large N) becomes inevitable.

(21.4) **The Jarzynski and Crooks equalities**

For spontaneous processes, the second law of thermodynamics relates the entropy S to work W and free energy F by an inequality since the Boltzmann H-theorem proves that entropy must increase. Jarzynski and Crooks recently have discovered remarkable **equalities** for nonequilibrium thermodynamics that assume special importance for fluctuations that are observed in small systems.

(a) The Jarzynski equality between the work W and free energy change ΔF in a cyclic process is $\langle e^{-\frac{W}{k_bT}} \rangle = e^{-\frac{\Delta F}{k_bT}}$, where the brackets indicate an average performed over several cycles of an identical process, and thus includes the fluctuations in each cycle. Read about the equality, and note that its derivation uses the Liouville theorem. Comment on its importance in small systems.

(b) The Crooks equality is $\frac{\rho_F(W)}{\rho_R(W)} = e^{\frac{W-\Delta F}{k_b T}}$, where the $\rho_F(W)$ indicates the probability density of W, the work done in the forward path and $\rho_R(W)$ is the same for the reverse path. The spreads in $\rho(W)$ indicate the fluctuations in the forward and reverse processes. Read and appreciate this relation, and discuss its importance in nonequilibrium processes occurring in the nanoscale.

(21.5) **Putting the genie back in the bottle**
Newtons' laws of classical mechanics and Schrödinger's equation of quantum mechanics do not distinguish between processes running forward or backwards in time. In other words, these laws cannot describe why the molecules of perfume that escape from a vial into a room can never get back into the vial. This is explained by thermodynamics, which is not contained in Newton's and Schrödinger's equations. Describe how thermodynamics is introduced into classical mechanics (resulting in Maxwell–Boltzmann distributions) and into quantum mechanics (resulting in Fermi–Dirac and Bose–Einstein distributions), and how this maintains the irreversibility of certain processes.

(21.6) **Legacy of the BTE: 150 years of heated debates, beautiful physics and mathematics**
Famously, Poincare proved a recurrence theorem which shows that the gas escaping from a bottle actually *will* return into the bottle, but after an astronomically long time, much longer than the age of the universe! The H-theorem has a rich legacy: it introduced a quantitative formulation of entropy. This in turn led Planck to realize that since the entropy in the presence of a continuous energy spectrum of photons would become infinite, the only resolution was to assume a discrete, or quanta of photon energies. This realization gave birth to quantum mechanics. The mathematics underlying the H-theorem treatment of the BTE has been a rich source of ideas for partial differential equations, leading to several fundamental breakthroughs. The latest and most famous is the proof of the century old Poincare conjecture by G. Perelman in 2006, who among other things used a concept (similar to Boltzmann's H-entropy) of the W-entropy of solitons flowing along Ricci fields. This is ironic given Poincare's recurrence theorem's role against Boltzmann's H-theorem! Study this chain of events and write a short essay based on the connections created by the BTE.

(21.7) **The Maxwell–Boltzmann distribution from the BTE**
As an example of the original result derived by Boltzmann from the BTE, using the entropy function for classical particles in Equation 4.5, derive the Maxwell–Boltzmann distribution for particles with mass M and velocities v using the same strategy that was used to derive the Fermi–Dirac distribution in Section 21.5. The result is $f(v) = (\frac{m}{2\pi k_b T})^{\frac{3}{2}} \cdot 4\pi v^2 \cdot e^{-\frac{mv^2}{2k_b T}}$, which is proportional to $v^2 e^{-A \cdot v^2}$. Sketch the distribution in velocities at several temperatures.

(21.8) **Bose–Einstein Distribution from the BTE**
As another example of a rather remarkable result derivable from the BTE, using the entropy function for bosons in Equation 4.4, derive the Bose–Einstein distribution using the same strategy that was used to derive the Fermi–Dirac distribution in Section 21.5.

(21.9) **The electro-magneto-thermal vector G**
The generalized solution $f_{\mathbf{k}} = f^0_{\mathbf{k}} - \frac{\partial f^0_k}{\partial E_{\mathbf{k}}} \phi_{\mathbf{k}}$ of the BTE in the simultaneous presence of electric and magnetic field and thermal gradient is given by $\phi_{\mathbf{k}} = -\tau_{\mathbf{k}}(\mathbf{v_k} \cdot \mathbf{G})$, where $\mathbf{G}$ is the electro-thermo-magnetic vector $\mathbf{G} = \frac{\mathbf{F} + q\tau(\mathbf{B} \times [M]^{-1}\mathbf{F}) + \frac{(q\tau)^2}{\det[M]}(\mathbf{F} \cdot \mathbf{B})[M]\mathbf{B}}{1 + \frac{(q\tau)^2}{\det[M]}(\mathbf{B} \cdot [M]\mathbf{B})}$, where $\mathbf{F} = \nabla_{\mathbf{r}}(\mu - q\varphi) + \frac{(E-\mu)}{T}\nabla_{\mathbf{r}} T$ is the electro-thermal force, $\mathbf{B}$ is the magnetic field, and $[M]$ is the effective mass tensor. In this problem we derive this from the BTE for scalar effective masses $m^\star$, and then investigate some properties of tensor effective masses.

(a) The collision term in the RHS of the BTE in the RTA is $\frac{\partial f_{\mathbf{k}}}{\partial t}|_{\text{coll.}} = -\frac{(f_{\mathbf{k}} - f^0_{\mathbf{k}})}{\tau_{\mathbf{k}}}$. The steady state drift terms in the LHS of the BTE with external field $\mathbf{F}_{\text{ext}} = -q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$ is $\frac{\mathbf{F}_{\text{ext}}}{\hbar} \cdot \nabla_{\mathbf{k}} f_{\mathbf{k}} + \mathbf{v_k} \cdot \nabla_{\mathbf{r}} f_{\mathbf{k}}$. Show by using $f_{\mathbf{k}} = f^0_{\mathbf{k}} - \frac{\partial f^0_k}{\partial E_{\mathbf{k}}} \phi_{\mathbf{k}}$, substituting the electro-thermal force $\mathbf{F}$ and the external force $\mathbf{F}_{\text{ext}}$, and for a small change in $f_{\mathbf{k}}$, the BTE becomes $-(\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})\mathbf{v_k} \cdot \mathbf{F} + (\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})\frac{q}{\hbar}(\mathbf{v_k} \times \mathbf{B}) \cdot \nabla_{\mathbf{k}} \phi_{\mathbf{k}} = (\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})\frac{\phi_{\mathbf{k}}}{\tau_{\mathbf{k}}}$.

(b) Cancelling $(\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})$ and rearranging, the vector differential equation $\phi_{\mathbf{k}} = -\tau_{\mathbf{k}}(\mathbf{v_k} \cdot \mathbf{F}) + \frac{q\tau_{\mathbf{k}}}{\hbar}(\mathbf{v_k} \times \mathbf{B}) \cdot (\nabla_{\mathbf{k}} \phi_{\mathbf{k}})$ needs to be solved for $\phi_{\mathbf{k}}$. Because the solution for $\mathbf{B} = 0$ is $\phi_{\mathbf{k}} = -\tau_{\mathbf{k}}(\mathbf{v_k} \cdot \mathbf{F})$, try a solution of the form $\phi_{\mathbf{k}} = -\tau_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}} \cdot \mathbf{G})$, where $\mathbf{G}$ should be a vector composed of both $\mathbf{B}$ and $\mathbf{F}$. Show that for a scalar effective mass, the form in Equation 21.52 is indeed obtained: $\mathbf{G} = \frac{\mathbf{F}+\mu(\mathbf{B}\times\mathbf{F})+\mu^2(\mathbf{F}\cdot\mathbf{B})\mathbf{B}}{1+\mu^2|\mathbf{B}|^2}$.

Solution: For a scalar effective mass $\nabla_{\mathbf{k}} = \frac{\hbar}{m^\star}\nabla_{\mathbf{v}}$ gives $\nabla_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}} \cdot \mathbf{G}) = \frac{\hbar}{m^\star}\nabla_{\mathbf{v}}(\mathbf{v} \cdot \mathbf{G})$, where the subscript $\mathbf{k}$ is suppressed. In the resulting identity $\nabla_{\mathbf{v}}(\mathbf{v} \cdot \mathbf{G}) = (\mathbf{G} \cdot \nabla_{\mathbf{v}})\mathbf{v} + (\mathbf{v} \cdot \nabla_{\mathbf{v}})\mathbf{G} + \mathbf{v} \times (\nabla_{\mathbf{v}} \times \mathbf{G}) + \mathbf{G} \times (\nabla_{\mathbf{v}} \times \mathbf{v})$, the last three terms are zero since $\nabla_{\mathbf{v}} \times \mathbf{v} = 0$ and $\mathbf{G}$ does not depend physically on $\mathbf{v}$. Only the first term survives, which upon writing out the components becomes simply $(\mathbf{G} \cdot \nabla_{\mathbf{v}})\mathbf{v} = [G_1, G_2, G_3][\partial_{v_1}, \partial_{v_2}, \partial_{v_3}]^T[v_1, v_2, v_2] = [G_1, G_2, G_3] = \mathbf{G}$. Substituting $\phi_{\mathbf{k}} = -\tau_{\mathbf{k}}(\mathbf{v}_{\mathbf{k}} \cdot \mathbf{G})$, the vector differential equation is reduced to the vector algebraic equation $\mathbf{v}_{\mathbf{k}} \cdot \mathbf{G} = \mathbf{v}_{\mathbf{k}} \cdot \mathbf{F} + \frac{q\tau_{\mathbf{k}}}{m^\star}(\mathbf{v}_{\mathbf{k}} \times \mathbf{B}) \cdot \mathbf{G}$. Using $\mu = q\tau_{\mathbf{k}}/m^\star$ and upon using the triple product property $(\mathbf{v}_{\mathbf{k}} \times \mathbf{B}) \cdot \mathbf{G} = \mathbf{v}_{\mathbf{k}} \cdot (\mathbf{B} \times \mathbf{G})$, the group velocity $\mathbf{v}_{\mathbf{k}}$ cancels, giving $\mathbf{G} = \mathbf{F} + \mu(\mathbf{B} \times \mathbf{G})$. This equation is solved for $\mathbf{G}$ by taking a dot product with $\mathbf{B}$ on both sides to give $\mathbf{G} \cdot \mathbf{B} = \mathbf{F} \cdot \mathbf{B}$ since $\mathbf{B} \cdot \mathbf{B} \times \mathbf{G} = 0$, and then taking the cross product on the left with $\mathbf{B}$ to give $\mathbf{B} \times \mathbf{G} = \mathbf{B} \times \mathbf{F} + \mu\mathbf{B} \times (\mathbf{B} \times \mathbf{G}) = \mathbf{B} \times \mathbf{F} + \mu[(\mathbf{B} \cdot \mathbf{G})\mathbf{B} - |\mathbf{B}|^2\mathbf{G}] = \mathbf{B} \times \mathbf{F} + \mu[(\mathbf{F} \cdot \mathbf{B})\mathbf{B} - |\mathbf{B}|^2\mathbf{G}]$, where the dot product result from above is used. The equation $\mathbf{G} = \mathbf{F} + \mu(\mathbf{B} \times \mathbf{F}) + \mu^2[(\mathbf{F} \cdot \mathbf{B})\mathbf{B} - |\mathbf{B}|^2\mathbf{G}]$ thus becomes linear, and upon rearrangement immediately gives $\mathbf{G} = \frac{\mathbf{F}+\mu(\mathbf{B}\times\mathbf{F})+\mu^2(\mathbf{F}\cdot\mathbf{B})\mathbf{B}}{1+\mu^2|\mathbf{B}|^2}$.

(c) Consider the changes above for the ellipsoidal conduction band of silicon say in the (100) direction: $E_{\mathbf{k}} = \frac{\hbar^2}{2}(\frac{k_x^2}{m_l} + \frac{k_y^2}{m_t} + \frac{k_z^2}{m_t})$, which is anisotropic. Show that the inverse effective mass tensor $[M]^{-1}$ of silicon is diagonal, and its determinant is $\frac{1}{\det[M]} = \frac{1}{m_l m_t^2}$. Use this in the expressions and find the electro-magneto-thermal vector $\mathbf{G}$ for conduction electrons in silicon. Discuss the relative strengths of the three terms: the conductivity, the Hall, and the magnetoresistance.

(d) Show that for tensor effective masses, the generalized kinetic coefficient can be expressed as $K_{rs}^{ij} = \frac{1}{L^d} \cdot (\frac{q}{m})^{r-1} \cdot \sum_{\mathbf{k}}(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}})\frac{v_i v_j \tau_{\mathbf{k}}^r (E_{\mathbf{k}}-\mu)^s}{1+\frac{q^2\tau^2}{\det[M]}(\mathbf{B}\cdot[M]\mathbf{B})}$ where the effective mass tensor matrix is $[M]$, and $\det[M]$ is its determinant.

(21.10) No change in net energy of drifting electrons

Show that in the relaxation-time approximation, the net internal energy of electrons drifting in response to an electric field $\mathbf{E}$ is the same as in equilibrium, when there was no field. Explain why this is the case physically by considering electrons moving along and against the electric field, and the fact that the energy $E_{\mathbf{k}}$ is even in $\mathbf{k}$.

(21.11) Thermoelectric coefficients in d-dimensions

(a) Using the fundamental kinetic coefficients of Equation 21.55, show that Seebeck coefficient for a parabolic bandstructure for transport in d-dimensions with a scattering time $\tau_{\mathbf{k}}$ is $S = -\frac{k_b}{q}[-\eta + \frac{\int_0^\infty du \frac{u^{\frac{d}{2}+1}\tau_u}{\cosh^2[\frac{u-\eta}{2}]}}{\int_0^\infty du \frac{u^{\frac{d}{2}}\tau_u}{\cosh^2[\frac{u-\eta}{2}]}}]$, where $\eta = \frac{\mu}{k_b T} = \frac{E_F - E_c}{k_b T}$.

Hint: Use $u = E_{\mathbf{k}}/k_b T$, the fact that the DOS for d-dimensions goes as $E_{\mathbf{k}}^{(d-2)/2}$, and notice that in the $(E_{\mathbf{k}} - \mu)$ term in the numerator, the μ is a constant for which the integrals in the numerator and the denominator cancel.

(b) The Seebeck coefficient depends on how the scattering time $\tau_{\mathbf{k}}$ depends on state $\mathbf{k}$, or equivalently the state's energy $E_{\mathbf{k}}$. In Chapters 22 and 23 it is shown that for several scattering mechanisms the scattering time can be expressed as a power law of $E_{\mathbf{k}}$ in the form: $\tau_{\mathbf{k}} = \tau_0(\frac{E_{\mathbf{k}}}{k_b T})^r = \tau_0 u^r$. Using this in part (a), show that the Seebeck coefficient simplifies to the forms given in Equation 21.73.

(c) Evaluate the Seebeck coefficient for various dimensions d and parameters r and show that they follow what is shown in Fig. 21.19. Argue why in the non-degenerate situation, $|S| \sim \frac{k_b}{q}|\eta|$ is an excellent approximation for the magnitude.

(21.12) Generalized d-dimensional Lorenz number

Show that the generalized d-dimensional Lorenz number using only the electronic part of the thermal conductivity for parabolic bands is $\mathcal{L} = (\frac{k_b}{q})^2[(\frac{d+2}{2} + r)(\frac{d+4}{2} + r)\frac{F_{\frac{d}{2}+r+1}(\eta)}{F_{\frac{d}{2}+r-1}(\eta)} - (\frac{d+2}{2} + r)^2(\frac{F_{\frac{d}{2}+r}(\eta)}{F_{\frac{d}{2}+r-1}(\eta)})^2]$, and that this reduces to Equation 21.24 in the non-degenerate and degenerate limits.

(21.13) Integration by parts

Show by using integration by parts that writing $u = E_{\mathbf{k}}/k_b T$ and $\eta = \mu/k_b T$ and the Fermi–Dirac distribution as $f(u) = \frac{1}{\exp[u-\eta]+1}$, the following identity holds: $\int_0^\infty(-\frac{\partial f(u)}{\partial u})G(u)du = \int_0^\infty du f(u)\frac{dG(u)}{du} - [f(u)G(u)]_0^\infty$. If $G(0) = 0$, then the product $f(u)G(u)$ vanishes at $u = 0$ and at $u =$

∞, leaving $\int_0^\infty (-\frac{\partial f(u)}{\partial u}) G(u) du = \int_0^\infty du f(u) \frac{dG(u)}{du}$, which maps to a sum of Fermi–Dirac integrals if $\frac{dG(u)}{du}$ is a polynomial in u. This trick is very useful in the evaluation of the kinetic coefficients of Equation 21.55.

(21.14) **Mott variable range hopping conductivity**

Defects in heavily disordered semiconductors introduce electron energy eigenvalues deep in the gap, far from the bands. An example is shown in Fig. 21.28. In such semiconductors, electrons are captured by the deep levels and are stuck there, occasionally to be re-emitted by thermal excitation. When an electric field is applied, the bands bend, and it becomes possible for an electron to hop from one deep level site to another. Such form of transport is called variable-range hopping, and was first explained by Mott. The electrical conductivity because of such a process comes out to be of the form $\sigma = \sigma_0 e^{-(\frac{T_0}{T})^{\frac{1}{d+1}}}$ in d-spatial dimensions. In this problem, you derive this result.

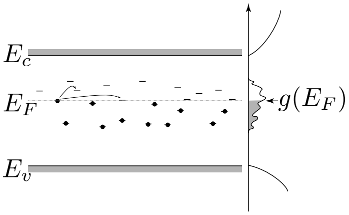


Fig. 21.28 Electron transport by hopping. The concept of variable range hopping was introduced by Mott.

(a) The current density $J = \sigma F$, where F is the electric field, and σ is the conductivity. The conductivity is conventionally thought of in the drift-picture as $\sigma = qn\mu$, where $\mu = q\tau/m^\star$ is the electron mobility. Instead of the band picture of the transport of electrons of effective mass $m^\star$ with a mean scattering time of τ, describe why the situation in the case of Fig. 21.28 is different. The electrons are localized and may not access the bands states: they may tunnel directly from one localized state to another. The localized wavefunctions vary as $\psi \sim e^{-r/\lambda}$ where λ is the spatial spread of the localized wavefunction and their populations for barrier height W go as the Boltzmann weight e^{-W/k_bT}. The hopping conductivity should be proportional to the product of these exponentials.

(b) Argue why the transition probability from site $i \to j$ is $S_{i\to j} = \frac{D_c^2 A_0^2 W}{\pi \rho_0 s^5 \hbar^4} \exp[-\frac{2R}{\lambda} - \frac{W}{k_bT}]$. Here D_c is the deformation potential (see Chapters 22 and 23), A_0 is an exchange integral, $W \approx k_b T_D = \hbar v_s q_D$ is the barrier height surmountable by phonon assisted tunneling where T_D is the Debye temperature and q_D the Debye wavevector, R is the distance between the localized sites, ρ_0 is the mass density of the semiconductor, v_s is the sound velocity, and T is the temperature.

(c) For $d = 3$, show that the extremum of W occurs at a radial distance $\bar{R} = (\frac{9\lambda}{8\pi k_b T g})^{\frac{1}{4}}$ and at the energy $\bar{W} = \frac{3}{4\pi \bar{R}^3 g}$, leading to the variable range conductivity of $\sigma = \sigma_0 e^{-(\frac{T_0}{T})^{\frac{1}{4}}}$, where $T_0 = \frac{5/2}{9\pi\lambda^3 k_b g}$. Here $g = g(E_F)$ is the DOS at the Fermi level.

(d) Consider the distance between localized states to be R, the attempt frequency for tunneling to be ν_{ph} to phonon frequency, tunneling decay length λ. Write the expression for the current density in terms of these parameters.

Solution: The current density is written in analogy to the formulation $J = qnv$. Identifying the terms, we get

$$\begin{aligned} J &\approx q \cdot g(E) k_b T \cdot R \nu_{ph} e^{-\frac{2R}{\lambda}} \cdot [e^{\frac{-\Delta E + qFR}{k_bT}} - e^{\frac{-\Delta E - qFR}{k_bT}}] \\ &\approx 2q \cdot g(E) k_b T \cdot R \nu_{ph} e^{-\frac{2R}{\lambda} - \frac{\Delta E}{k_bT}} \cdot \sinh[\frac{qFR}{k_bT}] \\ &\approx 2q^2 \cdot g(E) \cdot R^2 \nu_{ph} e^{-\frac{2R}{\lambda} - \frac{\Delta E}{k_bT}} \cdot F \\ \implies \sigma &= J/F \approx 2q^2 g(E) \nu_{ph} R^2 e^{-\frac{2R}{\lambda} - \frac{\Delta E}{k_bT}}. \end{aligned} \tag{21.85}$$

(e) Mott (Fig. 14.3) argued that instead of an electron tunneling to a nearest neighbor energy state that would be high in energy, the electron could tunnel further in space to states closer to it in energy. The number of such states available per unit energy interval are $\frac{4}{3}\pi R^3 \cdot g(E_F)$, which implies that the energy separation is $\Delta E = \frac{3}{4\pi R^3 g(E_F)}$. Show that this leads directly to the Mott variable range formula for hopping conductivity.

Solution: The maximum is reached for R such that

$$\frac{d}{dR}[-\frac{2R}{\lambda}-\frac{3}{4\pi R^3 g(E_F)k_bT}]=0,$$

$$\implies R_0^4=\frac{9\lambda}{8\pi g(E_F)k_bT} \text{ and } T_0=\frac{5/2}{9\pi k_b g(E_F)\lambda^3}$$

$$\boxed{\sigma = 3\nu_{ph}q^2\cdot\sqrt{\frac{\lambda g(E_F)}{2\pi k_bT}}\cdot e^{-(T_0/T)^{\frac{1}{4}}}.} \tag{21.86}$$

(f) Using similar arguments as above, now show that in d-dimensions, Mott's variable range hopping conductivity is given by the relation $\sigma = \sigma_0 e^{-(\frac{T_0}{T})^{\frac{1}{d+1}}}$, with σ_0 and T_0 specific to each dimension.

(21.15) Obtaining the dynamical form of the BTE

Provide the steps to arrive at the useful dynamical form of the BTE in Equation 21.83 from Equation 21.82. The two terms on the RHS of Equation 21.82 are $\frac{q}{\hbar\Omega}\sum_{\mathbf{k}}\mathbf{v_k}(\mathbf{F}\cdot\nabla_{\mathbf{k}}f_{\mathbf{k}})+\frac{q}{\Omega}\sum_{\mathbf{k}}\mathbf{v_k}(\mathbf{v_k}\cdot\nabla_{\mathbf{r}}f_{\mathbf{k}})$, where $\mathbf{F}=-q(\mathbf{E}+\mathbf{v_k}\times\mathbf{B})$ is the Lorentz force on the electron. Show that $\frac{q}{\hbar\Omega}\sum_{\mathbf{k}}\mathbf{v_k}(\mathbf{F}\cdot\nabla_{\mathbf{k}}f_{\mathbf{k}}) = \frac{nq^2}{m^\star}\mathbf{F}-\frac{q}{m^\star}(\mathbf{J}\times\mathbf{B})$ and $\frac{q}{\Omega}\sum_{\mathbf{k}}\mathbf{v_k}(\mathbf{v_k}\cdot\nabla_{\mathbf{r}}f_{\mathbf{k}})=+\frac{qD}{\tau_0}\nabla_{\mathbf{r}}n$, which results in the form given in Equation 21.83.

(21.16) Monte Carlo and path integrals

The Monte Carlo technique is used for solving the BTE, specially for understanding transport at high electric fields. Write down a flowchart that explains how this stochastic method is applied to evaluate the response of an ensemble of electrons to external fields. Also investigate the interesting origin of the Monte Carlo algorithm: which features mathematicians Stanislaw Ulam and John von Neumann, gambling, and the building of the atomic bomb.

(21.17) Boltzmann transport: scattering and mobility

This problem provides a preview of transport properties that will be the subject of Chapter 23.

(a) We derived the solution to the Boltzmann transport equation in the relaxation-time approximation for elastic scattering events to be $f(\mathbf{k})\approx f_0(\mathbf{k})+\tau(\mathbf{k})(-\frac{\partial f_0(\mathbf{k})}{\partial\mathcal{E}(\mathbf{k})})\mathbf{v_k}\cdot\mathbf{F}$, where all symbols have their usual meanings. Use this to show that for transport in d dimensions in response to a constant electric field E, in a semiconductor with an isotropic effective mass $m^\star$, the current density is $J=\frac{nq^2\langle\tau\rangle}{m^\star}E$, where $\langle\tau\rangle=\frac{2}{d}\cdot\frac{\int d\mathcal{E}\cdot\tau(\mathcal{E})\mathcal{E}^{\frac{d}{2}}(-\frac{\partial f_0(\mathcal{E})}{\partial\mathcal{E}})}{\int d\mathcal{E}\cdot\mathcal{E}^{\frac{d}{2}-1}f_0(\mathcal{E})}$, where the integration variable $\mathcal{E}=\mathcal{E}(\mathbf{k})$ is the kinetic energy of carriers. You have now at your disposal the *most general form* of conductivity and mobility from the Boltzmann equation for semiconductors that have a parabolic bandstructure! *Hint:* You may need the result that the volume of a d-dimensional sphere in the k-space is $V_d=\frac{\pi^{\frac{d}{2}}k^d}{\Gamma(\frac{d}{2}+1)}$, and some more dimensional and Γ-function information from Chapter 4.

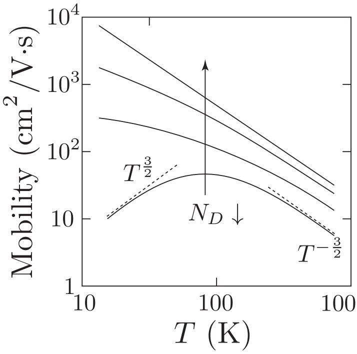


Fig. 21.29 Electron mobility in a doped semiconductor at high temperatures is limited by phonon scattering, and by impurity and defect scattering at low temperatures. In this problem, you show that the ionized-impurity scattering limited mobility goes as $T^{\frac{3}{2}}/N_D$.

(b) **Scattering from uncorrelated events**: Show using Fermi's golden rule that if the scattering rate of electrons in a band of a semiconductor due to the presence of *one* scatterer of potential $W(r)$ centered at the origin is $S(\mathbf{k}\to\mathbf{k}')=\frac{2\pi}{\hbar}|\langle\mathbf{k}'|W(r)|\mathbf{k}\rangle|^2\delta(E_{\mathbf{k}}-E_{\mathbf{k}'})$, then the scattering rate due to N_s scatterers distributed *randomly and uncorrelated* in 3D space is $N_s\cdot S(\mathbf{k}\to\mathbf{k}')$. In other words, the scattering rate increases *linearly* with the number of uncorrelated scatterers, which implies that the mobility limited by such scattering will decrease as $1/N_s$. This argument is subtle, and effects of electron wave interference should enter your analysis. *Hint*: Add the potentials of each randomly distributed impurity for the total potential $W_{tot}(r)=\sum_i W(\mathbf{r}-\mathbf{R_i})$. Use the effective mass equation for the electron states to show that the matrix element is a Fourier transform. Then invoke the shifting property of Fourier transforms.

(c) **Impurity scattering**: Using Fermi's golden rule, calculate the scattering rate for electrons due to a screened Coulombic charged impurity potential $V(r) = -\frac{Ze^2}{4\pi\epsilon_s r} e^{-\frac{r}{L_D}}$, where Ze is the charge of the impurity, ϵ_s is the dielectric constant of the semiconductor, and $L_D = \sqrt{\frac{\epsilon_s k_b T}{ne^2}}$ is the Debye screening length and n is the free carrier density. This is the scattering rate for just one impurity. Show using the result in parts (a) and (b), with a $1 - \cos\theta$ angular factor for mobility that if the charged-impurity density is N_D, the mobility for 3D carriers is $\mu_I = \frac{2^{\frac{7}{2}} (4\pi\epsilon_s)^2 (k_b T)^{\frac{3}{2}}}{\pi^{\frac{3}{2}} Z^2 e^3 \sqrt{m^\star} N_D F(\beta)} \sim \frac{T^{\frac{3}{2}}}{N_D}$. Here $\beta = 2\sqrt{\frac{2m^\star (3k_b T)}{\hbar^2}} L_D$ is a dimensionless parameter, and $F(\beta) = \ln[1 + \beta^2] - \frac{\beta^2}{1+\beta^2}$ is a weakly varying function. This famous result is named after Brooks and Herring who derived it first. Estimate the ionized impurity scattering limited mobility at $T = 300$ K for $m^\star = 0.2m_0$, $\epsilon_s = 10\epsilon_0$, $Z = 1$, and $N_D \sim 10^{17}, 10^{18}, 10^{19}/\text{cm}^3$. Are your values close to what is experimentally observed for these conditions as shown in Fig. 21.29?

We now explain the complete temperature dependence of the electron mobility in some (not all!) doped 3D semiconductors. Fig. 21.29 shows the experimental result: at low temperatures, the electron mobility *increases* with temperature as $\mu(T) \sim T^{\frac{3}{2}}/N_D$, and at high temperature it *decreases* with temperature as $\mu(T) \sim 1/T^{3/2}$. We first connect the mobility to the scattering times via the Drude-like result $\mu = \frac{q\langle\tau\rangle}{m_c^\star}$ where you found how to calculate the ensemble averaged scattering time $\langle\tau\rangle$ in part (a).

(d) **Phonon scattering**: The scattering rate of electrons due to acoustic phonons in semiconductors is given by Fermis golden rule result for time-dependent oscillating perturbations $\frac{1}{\tau(\mathbf{k}\to\mathbf{k}')} = \frac{2\pi}{\hbar} |\langle\mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle|^2 \delta(E_{\mathbf{k}} - E_{\mathbf{k}'} \pm \hbar\omega_q)$, where the acoustic phonon dispersion for low energy (or long wavelength) is $\omega_q \sim v_s q$ with v_s the sound velocity, and the scattering potential is $W(\mathbf{r}) = D_c \nabla_{\mathbf{r}} \cdot \mathbf{u}(\mathbf{r})$. Here D_c is the deformation potential (units: eV), and $\mathbf{u}(\mathbf{r}) = \hat{\mathbf{n}} u_0 e^{i\mathbf{q}\cdot\mathbf{r}}$ is the spatial part of the phonon displacement wave, $\hat{\mathbf{n}}$ is the unit vector in the direction of atomic vibration, and the phonon wavevector $\mathbf{q}$ points in the direction of the phonon wave propagation. We also justified why the amplitude of vibration u_0 may be found from $2M\omega_q^2 u_0^2 \approx N_{ph} \times \hbar\omega_q$, where $N_{ph} = 1/[e^{\frac{\hbar\omega_q}{k_b T}} - 1]$ is the Bose-number of phonons, and the mass of a unit cell of volume Ω is $M = \rho\Omega$, where ρ is the mass density (units: kg.m^{-3}). Show that a transverse acoustic (TA) phonon does *not* scatter electrons, but longitudinal acoustic (LA) phonons do.

(e) By evaluating the scattering rate using Fermi's golden rule, and using the ensemble averaging of part (a), show that the electron mobility in three dimensions due to LA phonon scattering is $\mu_{LA} = \frac{2\sqrt{2\pi}}{3} \frac{q\hbar^4 \rho v_s^2}{(m_c^\star)^{\frac{5}{2}} D_c^2 (k_b T)^{\frac{3}{2}}} \sim T^{-\frac{3}{2}}$. This is a *very* useful result.

(f) Now combine your work from all parts of this problem to explain the experimental dependence of mobility vs. temperature and as a function of impurity density as seen in Fig. 21.29.

(21.18) **The density-matrix formalism for transport**

The density matrix formalism of transport is based on the partition function which for a quantum system takes the form of the density matrix

$$\hat{\rho} = \frac{1}{\mathcal{Z}} \exp[-\beta\hat{H}], \tag{21.87}$$

whose time evolution is given by

$$i\hbar\frac{\partial\hat{\rho}}{\partial t} = [\hat{H}, \hat{\rho}], \tag{21.88}$$

where $[A, B] = AB - BA$ is the commutator. The Kubo formalism of transport is based on this fully quantum mechanical approach. The system Hamiltonian is split between the unperturbed equilibrium state Hamiltonian $\hat{H}_0$ and a perturbation: $\hat{H} = \hat{H}_0 + \hat{W}$. Study this formalism and write a short essay on how this method can be reduced to several approximations, one of which is the Boltzmann approximation. A distinction can also be made between open and closed quantum systems: for an open system in which particle number or energy can vary, the time evolution includes generation/recombination terms of the kind $i\hbar\frac{\partial\hat{\rho}}{\partial t} = [\hat{H}, \hat{\rho}] \pm \Gamma\hat{\rho}$ where Γ has units of energy.

Taking the Heat: Phonons and Electron-Phonon Interactions

22

22.1 Phonon effects: A résumé 531
22.2 Phonon dispersions and DOS 533
22.3 Optical conductivity 538
22.4 Lyddane–Sachs–Teller equation 540
22.5 Acoustic wave devices 541
22.6 Thermal conductivity 542
22.7 Phonon number quantization 543
22.8 Electron-phonon interaction 544
22.9 Chapter summary section 550
Further reading 551
Exercises 551

The periodic arrangement of atoms in a semiconductor crystal and the chemical bonds connecting them is the atomic-scale realization of a classical mass-spring network (or an electrical *LC* network in the form of a transmission line). Collisions of an external mass with a mass-spring network excites vibrational waves by exchanging momentum and energy. In much the same way, sound waves, electrons, or photons moving inside a crystal can exchange momentum and energy strongly with the crystal by either losing energy to, or absorbing energy from, the lattice vibrations. The vibrations are quantized: a quantum of vibrational energy is called a phonon (from ”phono” or sound). Because phonons and their interactions with electrons and photons are integral to the physical properties of semiconductor materials and devices, we devote this chapter to learn:

- What are the effects of phonons on the properties of solids in general, and semiconductors in particular?
- What are the allowed modes, energy dispersions, density of states, and quantum statistics of phonons in semiconductors?
- How do phonons influence the optical, electronic, and thermal conductivity of semiconductors?
- Which device applications of semiconductors are rooted in the properties of phonons?

22.1 Phonon effects: A résumé

In late 1890s a curious property of crystalline insulators called the **Reststrahlen band** was observed. When broadband infrared light was reflected off of the surface of the crystal, a narrow band of wavelengths reflected with near unity coefficient, as if there was a ”silver coating” only for this band. The specific wavelength range was characteristic of the crystal. Multiple reflections purified the infrared spectra further; the technique of using the Reststrahlen (”residual rays”) was instrumental in the development of quantum mechanics itself (Fig. 22.1). As

Fig. 22.1 Around 1900, Heinrich Rubens (above) discovered the Reststrahlen effect in the optical reflectivity of crystals working with Ernest Nichols. Using this Rubens constructed the Reststrahlen infrared spectrometer, with which he accurately measured the blackbody radiation spectrum. He shared his data with Max Planck, and the next day Planck had obtained his famous radiation formula which launched quantum mechanics.

we will see in this and later chapters, this narrow band of frequencies lies between the transverse optical (TO) and longitudinal optical (LO) branches of the phonon spectrum of a crystal[1].

An unresolved mystery in early 1900s was the temperature dependence of the **heat capacity** of solids, as we discussed in Chapter 4, specifically in Exercises 4.7, 4.8, and 4.9. By direct application of the classical laws of thermodynamics with the assumption of the presence of atoms, the Dulong–Petit law explained why the heat capacity per unit volume of a 3D solid is $c_v = 3nk_B$, where n is the atomic density and k_B the Boltzmann constant. Experimentally this $T-$independence is violated dramatically at low temperatures, since $c_v \to 0$ as $T \to 0$.

The first explanation for this behavior was offered by Einstein in 1907. Drawing on the analogy with the quantization of photons, he hypothesized that **lattice vibrations are quantized**. His bold claim was that the same rule of quantization of energy that applied to oscillating electromagnetic waves in blackbody radiation also applied to matter waves in solids. His simplified model assumed that all phonons have the same energy ω_0, and that phonons follow the **Bose–Einstein occupation function** in energy. We now know that this is the correct model only for **optical phonon** modes whose non-zero energy is indeed nearly constant, independent of wavelength.

Fig. 22.2 Peter Debye made important contributions to several areas of physics such as X-ray scattering and the heat capacity of solids. The Einstein and the Debye models of heat capacity are the earliest quantum theories of solids. Debye was awarded the 1936 Nobel Prize in Chemistry.

Though Einstein's model of the heat capacity of solids (the first application of quantum theory to solids) explained the vanishing of c_v as $T \to 0$ K, it was inconsistent with the quantitative temperature dependence, which went as $c_v \sim T^3$. This was fixed by Debye (Fig. 22.2) in 1912, who hypothesized the existence of a range of phonon frequencies ω_β starting from zero and increasing with the phonon wavevector $\beta(= 2\pi/\lambda)$ as $\omega_\beta = v_s\beta$, where v_s is the velocity of sound in the solid. Debye thus first hypothesized the existence of **acoustic phonon** modes in a solid: his model remains highly successful in explaining thermal properties of solids because the low energy acoustic phonon modes are occupied much more than the higher energy optical phonon modes.

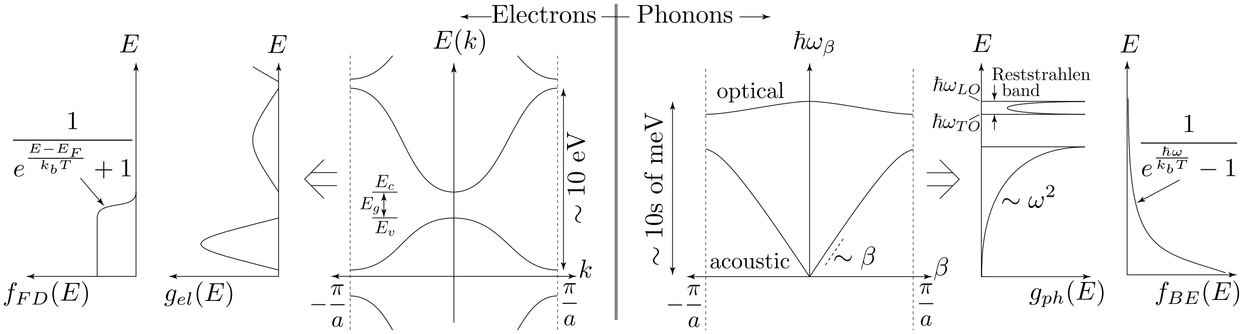
The phonons therefore have a "bandstructure" similar to that of electrons. They are measured experimentally by several complementary techniques. For very long wavelengths in the $q \to 0$ limit, **Raman spectroscopy** directly measures the zone-center optical phonon energies, and **Brillouin spectroscopy** measures the acoustic modes. **Neutron diffraction** and **inelastic X-ray scattering** techniques are used to measure the phonon dispersion across the entire Brillouin zone[2].

In semiconductors, under ordinary conditions, phonons influence carrier transport and optical phenomena to **perturbative order**. This means phonons do not change the electronic bandstructure, and only influence the transition rates between states via scattering, treated via Fermi's golden rule. But they dominate the propagation of heat (thermal conductivity) and sound. Sound transport in semiconductors forms the basis of devices such as surface acoustic wave (**SAW**) and bulk acoustic wave (**BAW**) filters and resonators. Phonons strongly coupled to electrons form quasiparticles called **polarons**, and phonons

---

[1] Experimentally it was known well before the advent of quantum mechanics that insulating solids with light atoms and strong chemical bonds (e.g. diamond, SiC) are excellent conductors of sound, heat, and light (e.g. optically transparent), and poor conductors of electricity. They of course violated the classical Wiedemann–Franz law because of their low electrical conductivity. We have seen that the reason is not because they have fewer electrons per se, but because of energy bandgaps and fewer *free* electrons near the Fermi level.

[2] Theoretical models today can accurately calculate the experimentally measured phonon dispersions of bulk solids, and are developed enough to guide the design of phonon spectra in multilayer structures. Models for the calculation of the phonon dispersions of solids range from semi-classical mass-spring kinds (Born) through those based on stretching and bending of actual chemical bonds (Keating), to the modern first principles DFT methods with increasing computational cost and accuracy.



←Electrons
Phonons →
E
$\frac{1}{e^{\frac{E-E_F}{k_bT}}+1}$
$f_{FD}(E)$
E
$g_{el}(E)$
$\Leftarrow$
$E(k)$
E_c
E_g
E_v
~ 10 eV
k
$-\frac{\pi}{a}$
$\frac{\pi}{a}$
$\hbar\omega_\beta$
optical
$\sim$ 10s of meV
acoustic
$\sim\beta$
β
$-\frac{\pi}{a}$
$\frac{\pi}{a}$
$\Rightarrow$
E
$\hbar\omega_{LO}$
$\hbar\omega_{TO}$
Reststrahlen band
$\sim\omega^2$
$g_{ph}(E)$
E
$\frac{1}{e^{\frac{\hbar\omega}{k_bT}}-1}$
$f_{BE}(E)$

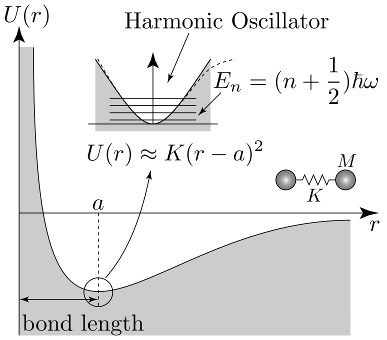
Fig. 22.3 A comparison of phonon dispersion, density of states, and occupation functions with the corresponding quantities for electrons in semiconductor crystals. Phonon energies are limited to $\sim$ 10s of meV, whereas electronic band energies range over $\sim$ 10 eV: note the large difference in scale! As opposed to electron energies, phonon energies are bounded from zero to a maximum optical phonon energy. Because of linear dispersion at low energies, the acoustic phonon DOS increases as the square of the energy (see Exercise 5.3). The Reststrahlen band lies between the LO and TO modes. As opposed to the Fermi–Dirac occupation function of electrons which are fermions, phonon occupation follows the Bose–Einstein distribution because phonons are bosons. Whereas electron occupation function is limited to a maximum of 1 due to the Pauli exclusion principle, no such limitation exists for phonons.

strongly coupled to photons form **polaritons**. A dramatic effect of strong electron-phonon coupling is the formation of Cooper pairs and **superconductivity**, which drastically changes the electronic properties into a new phase. Just as we developed the electronic energy bandstructures of semiconductors to understand their electronic and photonic properties, we begin our exploration of the physics of phonons in semiconductors by first investigating their energy bandstructure.

22.2 Phonon dispersions and DOS

Before describing the specific phonon dispersions of semiconductors and how they are measured or calculated, a direct comparison of their properties with those of electrons is both highly illustrative, and places the discussion in context. Fig. 22.3 shows them side by side. The properties of electrons on the left are discussed extensively earlier in this book. The corresponding properties of phonons are on the right. The figure label contrasts the properties, emphasizing the fact that phonons are bosons whose occupation function follows the Bose–Einstein statistics, and their energies are in the meV range for acoustic modes and in the 10s of meV range for optical modes.

We now discuss how the phonon dispersion (or bandstructure) of a crystal is obtained, starting from the vibrational properties of molecules. Fig. 22.4 shows the energy of interaction between two atoms when they are a distance r apart. For large r there is a weak van-der Waals attraction, whereas for very small r there is strong repulsion due to Coulomb interaction and Pauli exclusion. The distance of lowest energy $r = a$ is the equilibrium bond length. If the molecule is stretched from the equilibrium position $r = a$, the energy near the minimum



$U(r)$
Harmonic Oscillator
$E_n = (n+\frac{1}{2})\hbar\omega$
$U(r) \approx K(r-a)^2$
M
K
a
r
bond length

Fig. 22.4 Harmonic approximation. The interaction energy between atoms has a minimum, and near this minimum the energy is roughly quadratic in r. Small vibrations around the equilibrium position is modeled by a mass-spring system. The natural vibration frequency is $\omega = \sqrt{K/M_r}$, where $1/M_r = 1/M_1 + 1/M_2$ is the reduced mass of the two masses connected by a spring constant K. The lowering of energy drives the formation of H_2 or N_2 molecules.

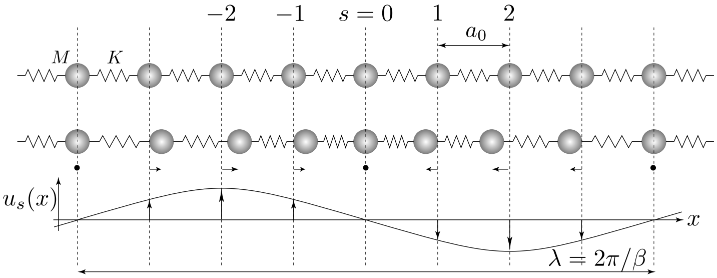


Fig. 22.5 Longitudinal acoustic (LA) phonons in a linear chain of atoms modeled as a mass-spring system. Note the adjacent atoms move in phase (i.e. in the same direction), a characteristic of acoustic phonons. Since the vibrations are along the x-axis which is the direction of the wave propagation, it is a longitudinal acoustic (LA) wave.

varies nearly quadratically with $r - a$, consistent with the harmonic oscillator approximation indicated in Fig. 22.4. Because the mass of the nuclei M are $\sim 10^3 \times m_e$ heavier than electron mass m_e, the vibrational energies are $\sim \frac{m_e}{M}$ times smaller for the same momentum. Because the electronic energy levels are in the 10s of eV, this pegs the energy scale of vibrations at 10s of meV. These vibrational energies fall in the infrared (IR) regime of the spectrum, and are directly measured by IR spectroscopy[3]. Because of the approximately parabolic potential, the energy levels are those of a quantum mechanical harmonic oscillator of energies $E_n = (n + \frac{1}{2})\hbar\omega$, where $\omega = \sqrt{2K/M}$ is the oscillation frequency, with K the spring constant and M the atomic mass. Lighter atoms and stronger chemical bonds lead to higher ω, but for a particular molecule, the frequency ω is fixed within the parabolic approximation. A higher quantized energy level n therefore implies a larger *amplitude* of oscillation at the same frequency.

Instead of two atoms, Fig. 22.5 shows their 1D chain of equilibrium lattice constant a_0. Imagine moving one atom away from its equilibrium position along x and letting it go. Since the atom is coupled to *all other* masses through the spring-mass network, this sets off a wave of vibration of *all* masses to travel through the chain in the x-direction. The energy imparted is saved in the system in the potential energies of the stretched or squeezed springs, and the kinetic energy of the moving masses. This is an example of a **collective excitation** in which energy is not stored in any particular atom, but collectively by all.

The harmonic approximation results in Hooke's law force $F = -K\Delta x$ on the atoms. If the displacement of the s^{th} atom from equilibrium is written as a propagating wave $u_s = u_0 e^{i(\beta x_s - \omega t)}$, nearest neighbor interactions on the s^{th} atom for $x_s = s a_0$ results in a restoring force[4]

$$F = M\frac{d^2 u_s}{dt^2} = K(u_{s+1} - u_s) + K(u_{s-1} - u_s), \text{ with } u_s(x) = u_0 e^{i(\beta s a_0 - \omega t)}$$

$$\implies -M\omega^2 = K(e^{i\beta a_0} - 1) + K(e^{-i\beta a_0} - 1)$$

$$\omega_\beta^2 = \frac{2K(1 - \cos\beta a_0)}{M}, \text{ and } \boxed{\omega_\beta = \sqrt{\frac{4K}{M}}\left|\sin\frac{\beta a_0}{2}\right| \approx \underbrace{\left[a_0\sqrt{\frac{K}{M}}\right]}_{v_s}\beta = v_s\beta,} \tag{22.1}$$

[3] Note that IR spectroscopy also forms the basis for gas detection and pollution control because the vibrational frequencies are like "fingerprints" of the specific molecular species. Semiconductor photonics is used for both the sources and detectors of IR photons for such studies, as described in Chapters 28 and 29.

[4] Note that though x is a continuous spatial variable, because $u_s(x)$ is the displacement of the s^{th} atom, it is *defined only at* the values $x_s = s a_0$ where $s = ... \pm 1, 0, 1, ...$ and not for any other x. Going from the first to second line of Equation 22.1, $u_0 e^{i(\beta s a_0 - \omega t)}$ cancels from both sides.

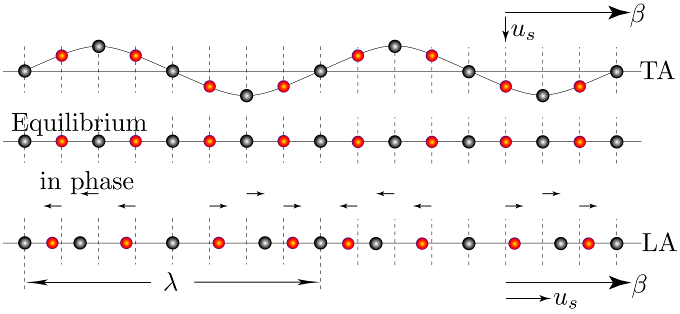


Fig. 22.6 Longitudinal Acoustic (LA) and transverse acoustic (TA) Phonons. The atoms vibrate parallel to the wave propagation direction $\vec{u}_s \parallel \vec{\beta}$ in the longitudinal modes and perpendicular $\vec{u}_s \perp \vec{\beta}$ to the wave propagation direction in the transverse modes. Note that the adjacent atoms may be different.

which yields an explicit dispersion relation of the phonon modes. For this phonon dispersion, the long-wavelength limit $\lambda \to \infty$ and $\beta = 2\pi/\lambda \to 0$ give $v_s = a_0\sqrt{K/M}$ with the dimensions of a velocity. This is the speed of sound propagation through the mass-spring system, and the modes are called "acoustic" phonons.

For periodic boundary conditions imposed on the chain (imagining it as a part of a large circular ring) of length L, the allowed wavevectors are $\beta_n = n \cdot \frac{2\pi}{L}$ where n is an integer, there are $N = L/a$ phonon modes that define a Brillouin zone $-\frac{\pi}{a} \leq \beta \leq +\frac{\pi}{a}$. The resulting acoustic mode dispersion $\hbar\omega_\beta = \hbar\sqrt{\frac{4K}{M}}|\sin\frac{\beta a_0}{2}|$ inside the BZ, and the resulting acoustic phonon density of states $g_{ph}(E)$ are shown in Fig. 22.3, with an analogy to the electron bandstructure and energy dispersion.

For a typical sound velocity $v_s \sim 10^5$ cm/s in a semiconductor, a phonon mode of wavelength $\lambda \sim 1$ μm has energy $\hbar v_s \beta \sim 20$ μeV. This energy is so small that even at room temperature of $T = 300$ K, $\hbar v_s \beta << k_b T$, and via the Bose–Einstein occupation function the number of vibration quanta in this mode corresponds to $N = 1/(e^{\hbar\omega_\beta/k_b T} - 1) \sim (k_b T)/(\hbar v_s \beta) \sim 1250$ acoustic phonons, a large number.

The presence of more than one atom in a unit cell leads to additional phonon mode branches. Specifically, **for N atoms in the unit cell there are $3N$ phonon branches or bands, of which 3 are acoustic branches and $3N-3$ are optical branches**. Let us discuss the additional acoustic branches and then move to optical branches. Consider Fig. 22.6 which shows a chain in which there are two atoms in the unit cell (they can be of the same mass). The lower part indicates a phonon wave in which the atomic displacement vector $\mathbf{u}_s$ is in the same direction x as the wave propagation $\vec{\beta}$: this is the **longitudinal acoustic** or LA mode. The atoms may also vibrate along y and z axes, which are independent modes and orthogonal to x. These two modes are called **transverse acoustic** or TA modes. The 1 LA and 2 TA modes for each β make up the 3 acoustic phonon branches. Using acoustic phonons in GaAs as an example, Fig. 22.7 shows why the LA phonon modes have higher energies than the TA modes. In a TA phonon wave moving in the [111] crystal direction, the atomic planes shear, or "slide against each other" laterally. In a LA mode, they are compressed or dilated. Because the

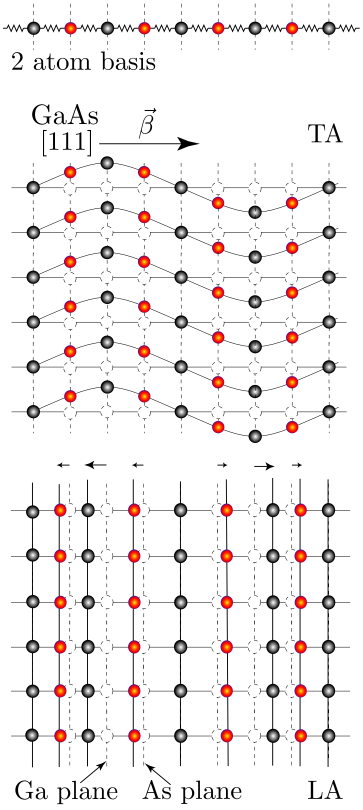


Fig. 22.7 Acoustic phonons propagating along the [111] direction in GaAs. The figure illustrates why the lattice strain caused by the LA phonon modes increases their energy to more than the TA modes, which primarily shear the crystal and do not strain it.

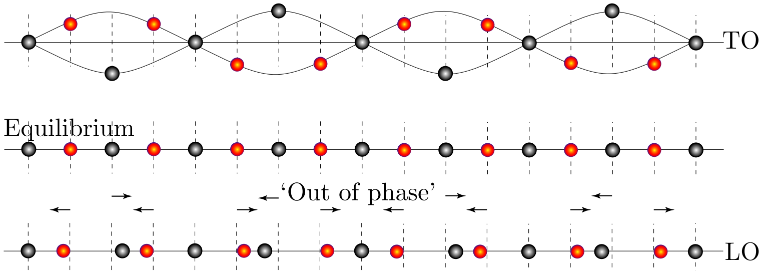


Fig. 22.8 Optical phonons, showing the vibration direction of nearest neighbors are out of phase – in opposite directions (compare with Fig. 22.6). The directions of vibration result in longitudinal optical (LO) and transverse optical (TO) phonons. Atoms in polar crystals carry effective charge, and result in charge dipoles leading to a macroscopic electric field. For $\beta \approx 0$ but $\beta \neq 0$, the LO energy is larger than the TO energy.

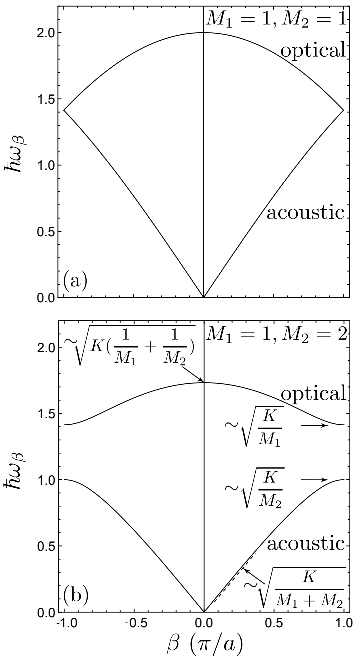


Fig. 22.9 Representative acoustic and optical phonon branches and their dependence on the atom masses M_1 and M_2. (a) Phonon dispersion for $M_1 = M_2$. If the two atoms have the same mass, the optical and acoustic branches touch as the BZ edge. (b) Phonon dispersion with $M_1 \neq M_2$. Heavier masses lower the phonon energies. Though a gap appears in the phonon spectrum in this 1D example of different masses, such gaps in the phonon spectra may or may not exist in real crystals of higher dimensions. Additional transverse acoustic (TA) and transverse (TO) modes are not captured in this toy model.

energy cost to compress or dilate the crystal is larger than to shear, the LA phonon branch energy is higher than the TA branch for all $\beta \neq 0$.

For $N > 1$ atoms in the unit cell, additional optical modes are formed. These vibrational modes are shown in Fig. 22.8. The major difference from the acoustic modes is that the vibrations of nearest neighbor atoms are out of phase, and in the opposite directions. They may be visualized as the combination of in-phase acoustic modes vibration of an atom of the unit cell which is 180° out of phase with that of the other atom of the unit cell. Writing the displacement of the two atoms as x_1 and x_2 and repeating the same procedure as Equation 22.1 for the 2 atom basis, the reader is encouraged to show that the new phonon energy dispersion for crystals with multi-atom basis is

$$\omega_\pm^2(\beta) = \frac{K}{M_r}[1 \pm \sqrt{1 - \frac{2M_r}{M_1 + M_2}(1 - \cos\beta a_0)}], \tag{22.2}$$

where $1/M_r = 1/M_1 + 1/M_2$ is the reduced mass. Solving this would require you to equate to zero a 2×2 determinant of the form $|D(\beta) - \omega_\beta^2 \delta_{ij}| = 0$. This toy model for phonons already has the ingredients for the calculation of accurate phonon dispersions. For the full phonon dispersion of solids, accuracy is increased by not limiting to nearest neighbor interactions, by solving a secular equation of a larger order dynamical vibration matrix with realistic "spring constants" K obtained from first principles. We do not pursue this calculation here, but discuss the results of the toy model captured in Equation 22.2 as shown in Fig. 22.9, and compare the result with Fig. 22.10 which shows the experimental phonon dispersions of Si, GaAs, and GaN, which are verified by accurate calculations.

Fig. 22.9 (a) shows that the ω_+ branch is the optical mode, and the ω_- branch is the acoustic mode which was discussed earlier. For semiconductors such as Ge, Si, or diamond, because the atoms in the basis have identical masses ($M_1 = M_2$), the branches touch at the Brillouin zone edge. Note the actual phonon dispersion of Si shown in Fig. 22.10 shows the LA and LO branches from the $\Gamma \rightarrow X$ points follow the same dispersion as Fig. 22.9 (a), including the fact that they touch at the X-point.

On the other hand, when the atoms in the unit cell are different,

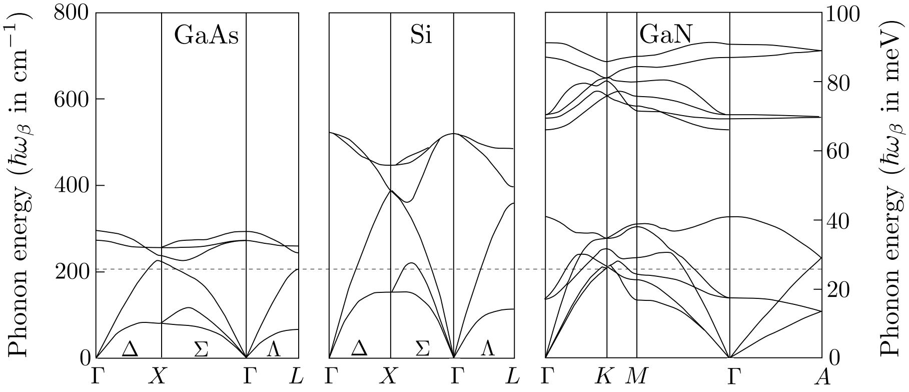


Fig. 22.10 Phonon dispersions of GaAs, silicon, and GaN. The phonon energy $\hbar\omega_\beta$ is shown on the left in wavenumber (cm^{-1}) units, and meV units on the right. The wavevectors β span paths of high symmetry in the Brillouin zone of the respective crystal structures (see Figures 10.10 and 10.12 for the paths). Note that the phonon energies increase for light atoms, and the substantial gap in the spectrum of GaN between optical and acoustic modes. The dashed line cutting across is $k_b T = 26$ meV at 300 K.

such as in GaAs, the LO and LA modes do not touch at the X-point as seen in both Fig. 22.9 (b) and Fig. 22.10 for GaAs. The maximum LA energy is proportional to $\sqrt{K/M_2}$ where M_2 is the mass of the heavier atom, and the LO energy at this wavevector is proportional to $\sqrt{K/M_1}$, where M_1 is the mass of the lighter atom. The optical phonon energies are dominated by the lighter atom due to their dependence on the *reduced* mass. Notice in Fig. 22.10 that the optical phonon energies increase substantially because at least one atom of the unit cell becomes lighter going from GaAs to Si to GaN, since the atoms are lighter. In the toy model, the optical phonon energy may be related roughly to the sound velocity via $\omega_{op} \approx 2 \cdot (v_s/a_0) \cdot (M_1 + M_2)/\sqrt{M_1 M_2}$.

The long wavelength ($\beta \to 0$) acoustic mode velocity is controlled by the *sum* of the masses, in which the heavy atom dominates. The sound velocity is higher for crystals made of lighter atoms and stronger chemical bonds. Compared to the LA phonon branches, the TA phonon dispersions of GaAs and Si start out and stay at lower energies both from $\Gamma \to \Delta \to X$ and $\Gamma \to \Lambda \to L$. The TA modes of GaAs and Si reach zero speed (or group velocity $v_g = \partial\omega_\beta/\partial\beta$) at the X and L points. The number of atoms per unit cell for GaAs and Si is $N = 2$, and that for GaN is $N = 4$. From Fig. 22.10, it is instructive to verify that the semiconductors indeed have $3N$ phonon branches, of which 3 are acoustic and $3N - 3$ are optical. The entire phonon spectra for all wavevectors β of most semiconductors can be measured by inelastic neutron spectroscopy (see Fig. 22.11). More recently, inelastic X-ray scattering has also been used for this purpose. The long-wavelength phonon energies near $\beta \approx 0$ however, can be measured in table-top systems by studying the reflection of light of much lower energy than X-rays. As discussed in Section 22.1, this was already achieved around 1900 before neutrons were discovered by studying the interaction of light with crystals, which is the topic of discussion in the next section.

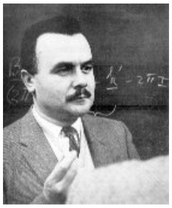


Fig. 22.11 B. Brockhouse (above) developed neutron scattering techniques to study solids with C. Shull, for which they shared the 1994 Nobel Prize in Physics. Brockhouse and P. K. Iyengar in 1958 measured the phonon dispersion of germanium. Nuclear reactors form the source of neutrons for such spectroscopy.

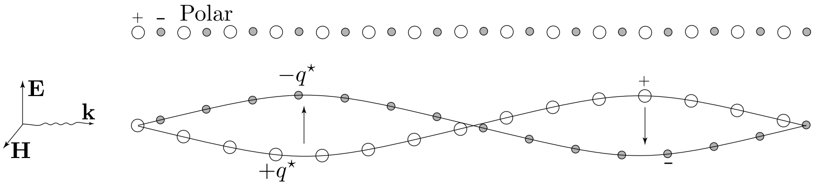


Fig. 22.12 Interaction of the TO phonon mode of a polar crystal with light in the form of a transverse electromagnetic wave.

22.3 Optical conductivity

The electric field component of light $E_0 e^{i(k \cdot x - \omega t)}$ interacts strongly with the transverse optical (TO) modes of polar semiconductors when the frequency of light matches that of the optical phonon $\omega \sim \omega_{TO}$. A simple reason is seen in Fig. 22.12. the polar crystal has atoms that have net alternating charge of opposite signs. Assuming light to be a transverse electromagnetic wave, the force due to the electric field of light pushes the positively and negatively charged atoms apart, generating a TO mode only if light is of the same wavelength as the TO phonon mode. The interaction of light with the polar semiconductor crystal is characterized quantitatively by a frequency dependent relative dielectric constant $\epsilon_r(\omega)$ shown in Fig. 22.13 (a), and a frequency-dependent reflectivity shown in Fig. 22.13 (b). We will now obtain the quantitative relationships using the classical **Drude–Lorentz oscillator model**. Let us assume that the net charge on the atom of mass M_+ is $+q^\star$ and of mass M_- is $-q^\star$. Let us also assume that the displacements of the atoms from their equilibrium position are x_+ and x_- respectively. We can then use Newton's law with a damping rate γ and $K = M\omega_{TO}^2$ to get for the two masses

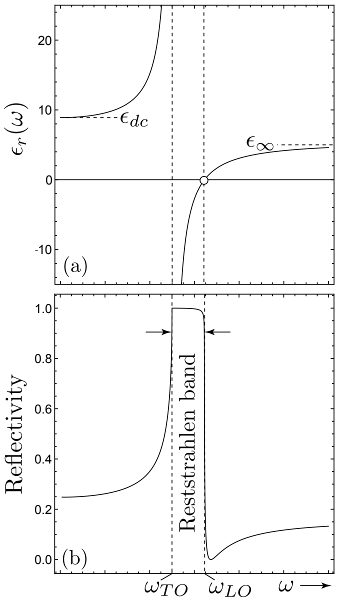


Fig. 22.13 For a polar semiconductor (e.g. GaN here), the frequency-dependent (a) dielectric constant $\epsilon_r(\omega)$ and (b) optical reflectivity are strongly influenced by the TO and the LO phonon modes.

$$\begin{aligned}
M_+ \frac{d^2 x_+}{dt^2} &= -M_+ \omega_{TO}^2 (x_+ - x_-) - M_+ \gamma \frac{dx_+}{dt} + q^\star E(t) \\
M_- \frac{d^2 x_-}{dt^2} &= -M_- \omega_{TO}^2 (x_- - x_+) - M_- \gamma \frac{dx_-}{dt} - q^\star E(t) \\
\implies \frac{d^2(x_+ - x_-)}{dt^2} + \omega_{TO}^2 (x_+ - x_-) + \gamma \frac{d(x_+ - x_-)}{dt} &= \frac{q^\star}{M_r} E(t),
\end{aligned} \tag{22.3}$$

and upon writing $x_+ - x_- = x_0 e^{i(\beta x - \omega_\beta t)}$ and $E(t) = E_0 e^{i(k \cdot x - \omega t)}$, it is seen that the modes only interact when the wavelength of light is matched to that of the phonon ($k = \beta$). This cancels the spatial dependence on both sides. Retaining the time dependence, we get

$$-\omega^2 x_0 + \omega_{TO}^2 x_0 - i\gamma x_0 = \frac{q^\star}{M_r} E_0 \implies x_0 = \frac{q^\star E_0 / M_r}{\omega_{TO}^2 - \omega^2 - i\gamma\omega} \tag{22.4}$$

as the amplitude of the relative vibration as a function of the electric field and the frequency of the lightwave[5]. Because $p = q^\star x_0$ is a microscopic dipole and there are N of them per unit volume, the electronic

[5] The damping, or "friction" term γ represents processes that scatter light, similar to the momentum scattering rate for electrons in the Drude model in Chapter 2. For example it could be dislocations in a crystal.

polarization that the external electric field of light produces is

$$P = Nq^\star x_0 = \frac{N(q^\star)^2}{M_r} \cdot \frac{1}{\omega_{TO}^2 - \omega^2 - i\gamma\omega} E_0. \tag{22.5}$$

$$\begin{gathered} \mathbf{D} = \epsilon_r \epsilon_0 \mathbf{E} = \epsilon_0 \mathbf{E} + \chi \epsilon_0 \mathbf{E} + \mathbf{P}_{ext} \implies \\ \epsilon_r(\omega) = 1 + \chi + \frac{N(q^\star)^2}{\epsilon_0 M_r} \cdot \frac{1}{\omega_{TO}^2 - \omega^2 - i\gamma\omega} \end{gathered} \tag{22.6}$$

which is an expression for the frequency-dependent relative dielectric constant, where χ is the electric suspectibility. Note that $\epsilon_r(\omega = 0) = 1 + \chi + \frac{N(q^\star)^2}{\epsilon_0 M_r \omega_{TO}^2} = \epsilon_{dc}$ is the low-frequency (dc) relative dielectric constant, and $\epsilon_r(\omega \to \infty) = 1 + \chi = \epsilon_\infty$ is the high-frequency relative dielectric constant. This implies $\epsilon_{dc} - \epsilon_\infty = \frac{N(q^\star)^2}{\epsilon_0 M_r \omega_{TO}^2}$, with which the frequency dependence of the relative dielectric constant is re-written in terms of experimentally measurable quantities as

$$\boxed{\epsilon_r(\omega) = \epsilon_\infty + (\epsilon_{dc} - \epsilon_\infty) \frac{\omega_{TO}^2}{\omega_{TO}^2 - \omega^2 - i\gamma\omega} = \epsilon_1 + i\epsilon_2.} \tag{22.7}$$

Table 22.1 Relation between real and imaginary parts of dielectric constants and refractive indices and the optical reflectivity.

$\eta + i\kappa = \sqrt{\epsilon_1 + i\epsilon_2}$
$\epsilon_1 = \eta^2 - \kappa^2$ and $\epsilon_2 = 2\eta\kappa$ $\eta = \sqrt{\frac{\sqrt{\epsilon_1^2+\epsilon_2^2}+\epsilon_1}{2}}$ and $\kappa = \sqrt{\frac{\sqrt{\epsilon_1^2+\epsilon_2^2}-\epsilon_1}{2}}$
$R = \frac{(\eta-1)^2+\kappa^2}{(\eta+1)^2+\kappa^2}$

Fig. 22.13 (a) shows an example of $\epsilon_r(\omega)$ for GaN with $\epsilon_{dc} = 8.9$, $\epsilon_\infty = 5.35$ and no damping ($\gamma = 0$). As the frequency of the light wave approaches the TO phonon frequency, the strong resonance sharply increases the relative dielectric constant from its DC value of ϵ_{dc}. $\epsilon_r(\omega)$ then is negative, and passes through zero at what will be seen in the next section to be the LO phonon frequency ω_{LO}, and increases towards ϵ_∞ at frequencies ω higher than ω_{LO}, but $\hbar\omega$ lower than interband transition energies. The entire frequency dependence of the optical response of semiconductors that include interactions in addition to phonons will be discussed in Chapter 27. The real and imaginary parts of the relative dielectric constant (ϵ_1, ϵ_2) are linked to the real and imaginary parts of the refractive index (η, κ) of the solid via the well-known relations given in Table 22.1, from which the optical reflectivity $R(\omega)$ is obtained using Equation 22.7 from experimentally known parameters. Fig. 22.13 (b) shows the calculated reflectivity corresponding to Fig. 22.13 (a) with a small nonzero γ. The semiconductor reflects IR light of frequencies between the TO and LO frequencies with $R \approx 1$: this is the **Reststrahlen band** discussed earlier. The reflectivity therefore constitutes a direct measurement of the TO and the LO phonon frequencies of a semiconductor. The deviation from $R \approx 1$ is a measure of the damping parameter γ.

The photonic bands are obtained by solving $\omega = \frac{c}{\sqrt{\epsilon_r(\omega)}} k$ with $\epsilon_r(\omega)$ from Equation 22.7. Of the four roots of ω obtained, two are positive. These branches are shown in Fig. 22.14 as a function of the light wavevector k. They are the allowed photon dispersion inside the semiconductor crystal, modified by the interaction with optical phonons.

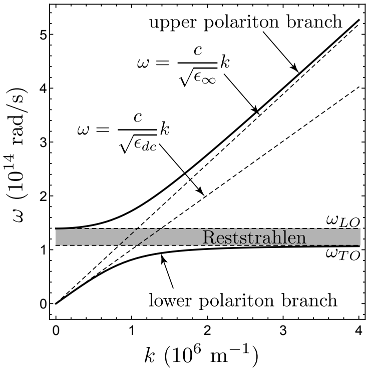


Fig. 22.14 Phonon-polariton modes for the propagation of light in the polar semiconductor with the simplified dielectric properties similar to GaN.

The strong interaction between photons and phonons splits the ordinary light dispersion into a upper and a lower branch. The modes near the intersection of the photon and phonon dispersions have the properties of both light and matter. These quasiparticles, called **phonon-polaritons**, have an energy gap in the Reststrahlen window.

22.4 Lyddane–Sachs–Teller equation

The relative dielectric constant vanishes at the LO phonon frequency, $\epsilon_r(\omega_{LO}) = 0$. An interesting relation between phonon frequencies and dielectric constants is obtained from Equation 22.7 (with $\gamma = 0$):

$$\boxed{\frac{\omega_{LO}^2}{\omega_{TO}^2} = \frac{\epsilon_{dc}}{\epsilon_\infty},} \tag{22.8}$$

which is called the **Lyddane–Sachs–Teller** equation after the researchers who derived it. Since light has very small $\mathbf{k}$, it interacts with the phonon modes near $\beta \approx 0$. Since for polar compound semiconductors $\omega_{LO} > \omega_{TO}$, ϵ_{dc} is larger than ϵ_∞. But for non-polar elemental semiconductors such as Si, Ge, or diamond, $\omega_{LO} = \omega_{TO}$ for $\beta \approx 0$, and therefore $\epsilon_{dc} = \epsilon_\infty$ in such semiconductors. They do not produce Reststrahlen bands, since they are not IR active, as is discussed briefly.

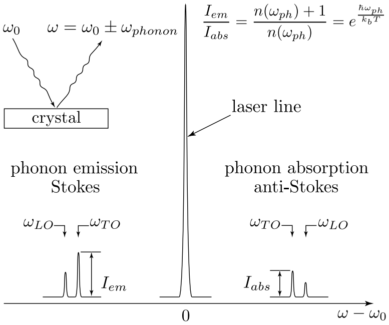


Fig. 22.15 Raman spectroscopy by inelastic scattering of light from crystals directly measures the Raman-active LO and TO modes. This method is used extensively in chemical identification and as a sensitive measurement of strain in semiconductor materials and heterostructures.

A remarkable consequence for the propagation of both light and electrons occurs as a result of the above considerations. Consider the requirement that in the absence of net unbalanced charge, the divergence of the displacement vector must vanish: $\nabla \cdot \mathbf{D} = 0$. Since $\mathbf{D} = \epsilon_0 \epsilon_r \mathbf{E}$, this implies $\epsilon_0 \epsilon_r(\omega)(\nabla \cdot \mathbf{E}) = 0$. Ordinarily, $\nabla \cdot \mathbf{E} = 0$ is ensured if $\nabla \cdot \mathbf{E} = \mathbf{E}_0 \nabla \cdot (e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}) = \mathbf{k} \cdot \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = 0$, which is true when $\mathbf{E}_0 \perp \mathbf{k}$, i.e., when the electric field $\mathbf{E}_0$ is perpendicular, or **transverse** to the propagation direction in which the wavevector $\mathbf{k}$ points. This is the case for a TEM wave. But for $\omega = \omega_{LO}$, a longitudinal EM wave with $\mathbf{E}_0 \parallel \mathbf{k}$ is sustained in a polar semiconductor crystal due to the formation of macroscopic electric fields due to the oscillating charged planes of atoms. The IR reflectivity drops to zero at ω slightly larger than ω_{LO} as seen in Fig. 22.13. We will discuss the implication on electron-phonon interaction in Section 22.8.

Fig. 22.16 C. V. Raman (above) with his student K. Krishnan discovered inelastic scattering of light as the visible wavelength analogue of X-ray Compton scattering. Raman was awarded the Nobel Prize in Physics in 1930.

When light is reflected off from a crystal, the photon can absorb or emit a phonon in the crystal. This inelastic scattering of light from crystals is called Raman scattering. It was discovered first for molecules, where light excites the molecular vibrational modes. The optical phonon vibrations of a semiconductor may be considered analogous to molecular vibrations, and similarly contain a "fingerprint" of the atomic constituents and also its strain state. Fig. 22.15 shows the principle of Raman spectroscopy, and Fig. 22.16 shows its discoveror.

For Raman spectroscopy of semiconductors, today a laser excitation with photon energy $\hbar\omega_0$ less than the semiconductor bandgap E_g is used. The reflected light has energies $\hbar\omega - \hbar\omega_{phonon}$ if a phonon is emitted by the light (called the Stokes process), and $\hbar\omega + \hbar\omega_{phonon}$ if a phonon is absorbed (called the anti-Stokes process). Since a phonon can only be absorbed if it is already there, the intensity of the anti-Stokes line for Raman-active TO or LO lines is $I_{abs} \propto n_{ph}$, where $n_{ph} = 1/(e^{\hbar\omega_{ph}/k_b T} - 1)$ is the Bose–Einstein occupation function. Thus, the intensity of the anti-Stokes line decreases with temperature. On the other hand, the intensity of the Stokes line for phonon emission is $I_{em} \propto 1 + n_{ph}$, the sum of the spontaneous and stimulated emission processes. Thus, the intensity ratio is $I_{abs}/I_{em} = e^{\hbar\omega_{ph}/k_b T}$, and the emission lines are stronger[6]. Raman spectroscopy involves three particles: two photons and one phonon, which makes it a weak process, requiring sensitive detectors. Fig. 22.17 shows experimental Raman spectra measured for single-crystal GaN and AlN.

22.5 Acoustic wave devices

The measurement of acoustic phonon modes by optical methods uses Brillouin scattering, which places even higher requirements on instrument sensitivity than Raman spectroscopy. On the other hand, electronic and microwave techniques are used to exploit sound wave propagation for devices whose importance in electronic communication systems is increasing rapidly. We discuss them in this section.

Acoustic phonon modes are responsible for the propagation of sound in insulating crystals. Typical sound velocities range from $v_s \sim 10^4 - 10^6$ cm/s in crystals, with semiconductors of lighter atoms exhibiting the highest velocities since $v_s \sim a_0\sqrt{K/(M_1 + M_2)}$ wherein the heavier mass dominates. By combining the toy model of phonons with the optical phonon energies, the longitudinal sound velocity near $\beta \approx 0$ may be approximated roughly by $v_L \sim a_L \omega_{LO}$, and the transverse sound velocity by $v_T = a_T \omega_{TO}$, where a_L and a_T are the respective lattice constants in the direction of wave propagation. We have already seen in Fig. 22.10 that the mode energies ω_β and as a result group velocities $\partial\omega_\beta/\partial\beta$ of the transverse acoustic branches are nearly half of the longitudinal branches. As a result, sound is dominantly carried by the longitudinal modes.

The longitudinal sound velocities are $v_s \sim 5.2 \times 10^5$ cm/s in GaAs, $\sim 8 \times 10^5$ cm/s in GaN, $\sim 11 \times 10^5$ cm/s in AlN, and $\sim 18 \times 10^5$ cm/s in diamond: note the increase in velocity with reduction in the mass of the atoms. Unlike diamond, GaAs, GaN, and AlN are **piezoelectric**. The combination of high acoustic sound velocities with piezoelectricity makes these crystals attractive for **electro-acoustic resonators** that are used as filters in RF and microwave electronics.

Such filters convert an electromagnetic wave of frequency f_0 into an acoustic wave of the same frequency. The wavelength shrinks in this process by the ratio of the sound velocity to the speed of light: $v_s/c \sim 10^{-4}$, which shrinks the filter size from several cm to microns, which is easily integrated on a chip. An incoming signal of frequency content

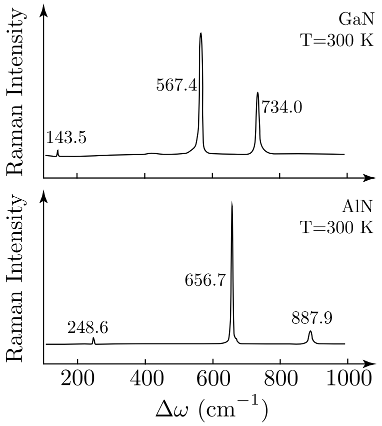


Fig. 22.17 Raman spectra measured in single-crystal GaN and AlN. Exercise 22.2 asks you to identify the optical phonon modes using Fig. 22.10, and to analyze the Lyddane–Sachs–Teller relation between the LO/TO phonon modes and the high/low frequency dielectric constants.

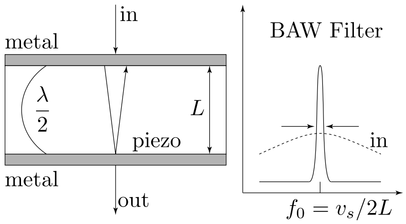


Fig. 22.18 Microwave filters commonly used in cellular phones use the bulk acoustic wave (BAW) resonator. AlN is widely used for BAW filters today because of the combination of desirable piezoelectric properties, high sound velocity for higher frequencies with the same size, and because it is compatible with fabrication methods. For example, $f_0 \sim 5.5$ GHz for AlN of $L = 1\ \mu$m. Section 22.8 discusses the physics of electromechanical coupling.

---

[6] Group theory is used to identify which optical phonon modes are Raman active, and which are IR active. As an example, the TO modes of GaAs are both Raman and IR active. The TO modes of silicon are Raman active, but not IR active.

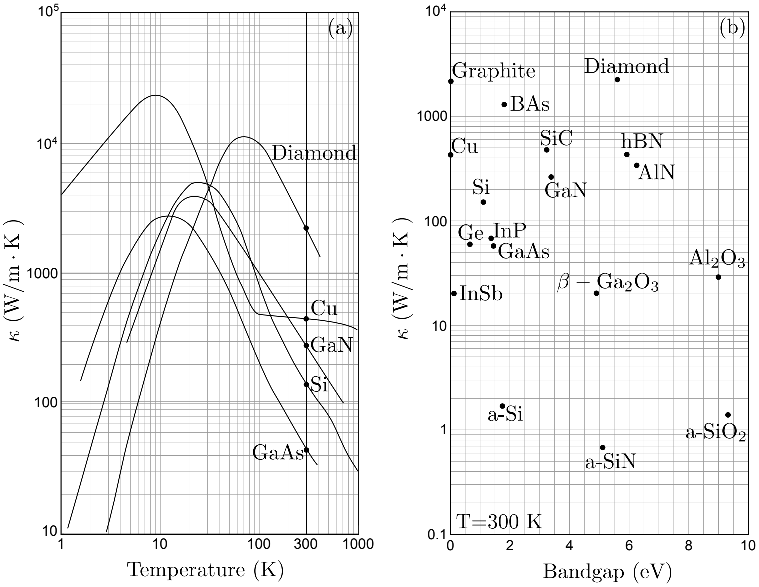


Fig. 22.19 Thermal conductivity of various semiconductors. (a) Thermal conductivity vs. temperature and with a comparison to metallic copper. (b) Room-temperature thermal conductivity as a function of the energy bandgap of various semiconductors, with those of typical insulators such as SiO_2 and SiN.

$f_0 + \Delta f$ is allowed to propagate through an acoustic wave cavity that only allows the desired signal of frequency f_0 and rejects the others. Such cavities may be formed with acoustic waves propagating along the surface (called surface-acoustic wave or **SAW**s), or bulk-acoustic wave or **BAW** resonators. Fig. 22.18 shows a BAW filter. A BAW cavity of length L filters the signal of frequency $f_0 \sim v_s/2L$ as the first mode for which half the acoustic wavelength fits in the cavity.

22.6 Thermal conductivity

Phonons carry heat in semiconductors. In Chapter 2, we saw how in metals the fact that heat energy is carried by the same electrons that carry electrical current led to the Wiedemann-Franz law. This law is violated for semiconductors, because in them electrical current is carried by electrons (and holes), but heat is carried primarily by lattice vibrations, or phonons. Metals such as copper are thought to be very good thermal conductors. But some of the best thermal conductors at 300 K are actually wide-bandgap semiconductors such as diamond and cubic BAs. SiC, AlN, and hBN approach the 300 K thermal conductivity of copper. Figures 22.19 (a) and (b) show the thermal conductivity of various materials.

The dependence of the thermal conductivity of semiconductors on temperature of Fig. 22.19 (a) has the following origin. Back in Chapter 2, Equation 2.7, we found that thermal conductivity is given by $\kappa = \frac{1}{3} c_v v^2 \tau$ where c_v is the specific heat capacity, v is the velocity of the heat carrier, and τ is the scattering time of the heat carrier. Since phonons are the heat carriers in semiconductors, their specific heat capacity follows the Debye law, going as $\sim T^3$ at the lowest temperatures, and saturating to the Dulong–Petit classical value of $\frac{3}{2}nk_b$ at temperatures higher than the Debye temperature Θ_D, which corresponds roughly to the highest acoustic phonon energy. The velocity is the sound velocity for long-wavelength acoustic modes, which is independent of temperature. The product $v\tau$ is an effective mean-free path, which at the lowest temperatures becomes equal to the sample size, which is independent of temperature.

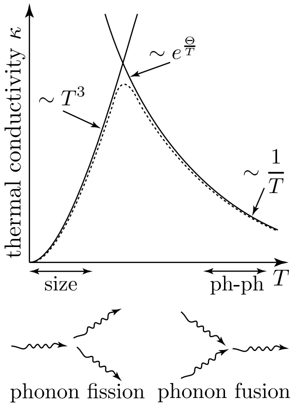


Fig. 22.20 Temperature dependence of thermal conductivity of semiconductors, and schematic representation of 3-phonon scattering processes due to the anharmonicity of the atomic potentials.

Therefore, the thermal conductivity at the lowest temperatures in ultrapure semiconductors with low defect densities is limited by **size-effect phonon scattering**, and goes as $\kappa \sim T^3$ as $T \to 0$ as shown in Fig. 22.20. Deviations from the T^3 low-temperature dependence typically are indications of the presence of defects and impurities that scatter phonons acoustically. Thus, κ increases sharply as the temperature increases. As the temperature approaches the Debye temperature Θ_D, **phonon-phonon scattering**, which was prohibited at the lowest temperatures by energy and momentum conservation requirements, becomes possible. This form of scattering is intrinsic as well, and arises from the anharmonicity of the crystal potential. The scattering rate by Fermi's golden rule is proportional to the density of phonons $1/\tau_{ph} \propto n_{ph} \propto 1/(e^{\Theta_D/T} - 1)$. Because at these temperatures c_v is T-independent, and v is always T-independent, the thermal conductivity $\kappa \propto (e^{\Theta_D/T} - 1)$ of semiconductors at high temperatures exhibits a maximum as shown in Fig. 22.20 schematically[7]. The corresponding experimental values at the highest temperatures drop as $\kappa \propto (e^{\Theta_D/T} - 1) \propto 1/T$. Exercise 22.4 outlines a more quantitative treatment of thermal conductivity and the effect of defects based on the BTE of Chapter 21.

22.7 Phonon number quantization

Let $N_\beta = 1/(e^{\hbar\omega_\beta/k_bT} - 1)$ be the Bose–Einstein occupation function for the phonon mode indexed by the wavevector β and of energy $\hbar\omega_\beta$ at temperature T. Consider a semiconductor of mass density ρ (e.g. in kg/m^3 units) and the volume of the entire crystal is L^3. For extensive use in the next section and in Chapter 23 to evaluate the scattering of electrons by phonons, we will now show the important quantum mechanical result that the squared vibration amplitude due to phonon absorption and phonon emission are given respectively by[8]

$$\boxed{|u_\beta^{abs}|^2 = \frac{\hbar N_\beta}{2\rho L^3 \omega_\beta}} \text{ and } \boxed{|u_\beta^{em}|^2 = \frac{\hbar (N_\beta + 1)}{2\rho L^3 \omega_\beta}}. \tag{22.9}$$

Each phonon mode is an independent harmonic oscillator. When phonons are in thermal equilibrium, the energy stored in the mode β of angular frequency ω_β is $E_\beta = (N_\beta + \frac{1}{2})\hbar\omega_\beta$, which is the sum of the kinetic energy of the moving atoms and the potential energy of

[7] The phonon-phonon scattering involves both fission and fusion processes because unlike electrons, their numbers need not be conserved. See Exercise 22.6 for a discussion why normal phonon processes are not enough, and Umklapp processes are needed to explain experimental thermal conductivities.

[8] We are drawing on a semi-classical version of the proof. The rigorous proof is provided in quantum mechanics textbooks for the harmonic oscillator.

the springs. The amplitude of vibrations of the standing wave and the corresponding kinetic energy stored in a unit volume are

$$u = u_\beta \cos \omega_\beta t \implies \mathrm{KE}(t) = 2\rho L^3 \omega_\beta^2 u_\beta^2 \sin^2 \omega_\beta t. \tag{22.10}$$

For an harmonic oscillator, $\langle \mathrm{KE} \rangle = \langle \mathrm{PE} \rangle$ over a period. With $\langle \sin^2 \omega t \rangle = 1/2$, the amplitude of vibrations of a given mode u_β in that case is[9]

$$\langle \mathrm{KE} \rangle + \langle \mathrm{PE} \rangle = 2\rho L^3 \omega_\beta^2 u_\beta^2 = (N_\beta + \frac{1}{2})\hbar\omega_\beta \implies u_\beta^2 = \frac{\hbar(N_\beta + \frac{1}{2})}{2\rho L^3 \omega_\beta}. \tag{22.11}$$

22.8 Electron-phonon interaction

In the preceding sections of this chapter we have discussed how the transport of light, heat, and sound through semiconductors is affected by phonons. This section is devoted to understand how phonons affect the transport of mobile electrons and holes. We saw that light either forms **quasiparticles** with phonons (e.g. phonon-polaritons) when the interaction is strong, or light is **scattered** by phonons when the interaction is weak (e.g. Raman or Brillouin scattering). Similarly, when the interaction of electrons with phonons is strong, they form quasiparticles such as polarons. Phonons simply scatter electrons when the interaction is weak. For most conventional semiconductors, the interaction is weak and the primary result of electron-phonon interactions is scattering. The strengths of the interactions are captured by writing the total energy of the electron-phonon system as

$$H = \underbrace{H_{\text{electron}}}_{E(k)} + \underbrace{H_{\text{phonon}}}_{\hbar\omega_\beta} + \underbrace{H_{\text{electron-phonon}}}_{W_\beta(r,t)}. \tag{22.12}$$

The first term is the electron energy in the presence of a perfectly still lattice (no phonons), which is the electron bandstructure $E(k)$. The second term allows the motion of the nuclei of atoms of the crystal around their equilibrium positions, and is the sum of the kinetic energy of the moving atoms and the potential energy of the "springs" between them: this is the phonon bandstructure $\hbar\omega_\beta$ discussed in this chapter. The third term is the energy of interaction of electrons and phonons. Fig. 22.21 shows a parabolic electron bandstructure $E(k)$ juxtaposed with the phonon bandstructure $\hbar\omega_\beta$ plotted for the same phonon and electron wavelengths with $\beta = k = 2\pi/\lambda$. These have already been discussed earlier. There are two intersections: one of electrons with optical phonons at $k_{op} \sim 0.1 \times \frac{\pi}{a}$, about 1/10th of the BZ, and another with acoustic phonons at $k_{ac} \sim 0.01 \times \frac{\pi}{a}$ at $\sim$ 1/100th of the BZ. Strong coupling may occur at these intersections leading to the formation of quasiparticles[10] – which are eigenvalues for the entire electron-phonon Hamiltonian of Equation 22.12. Since in most semiconductors the third term is much smaller than the first two, we will

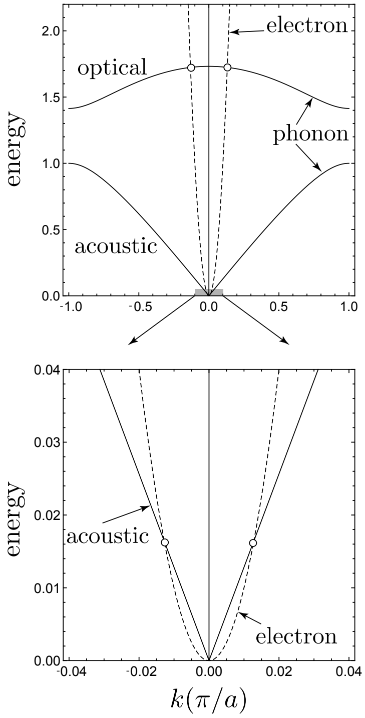


Fig. 22.21 A parabolic electron bandstructure $E(k)$ and the phonon bandstructure $\hbar\omega_k$ plotted to highlight the points of intersection.

---

[9] The squared amplitude for the phonon absorption process is proportional to the number of phonons present N_β (i.e. it is a *stimulated* process). But the squared amplitude for the emission process can be both stimulated and *spontaneous*, leading to the $N_\beta + 1$ dependence. This semi-classical argument thus gives the average of the two squared amplitudes of the exact quantum result of Equation 22.9.

[10] An extreme case of strong coupling of electrons and phonons results in the phenomena of superconductivity.

treat it as a perturbation $W_\beta(r,t)$ that depends on the electron wavevector k and phonon wavevector β. Typical energy scales are $E(k) \sim$ eV, $\hbar\omega_\beta \sim 1 - 10$s of meV, and $W_\beta(r,t) \sim 1$ meV or much less. The rest of this section is devoted to finding $W_\beta(r,t)$ in semiconductors.

Acoustic deformation potential (ADP): The displacement of an atom from its equilibrium position in a longitudinal acoustic (LA) wave is $\mathbf{u}_\beta(\mathbf{r}) = \hat{\beta}_L u_\beta e^{i(\vec{\beta}\cdot\mathbf{r}-\omega_\beta t)}$, where $\hat{\beta}_L$ is the unit vector along the longitudinal direction of the wave propagation, along the wavevector $\vec{\beta}$. The difference in displacement of atomic planes separated by a small distance $d\mathbf{r}$ is then $u_\beta(\mathbf{r}+d\mathbf{r}) - u_\beta(\mathbf{r}) \approx (\nabla \cdot \mathbf{u}_\beta(\mathbf{r})) \cdot d\mathbf{r}$. The spatial derivative or divergence[11] of the displacement $\mathbf{u}_\beta(\mathbf{r})$ is dimensionless, and is physically the "compression" or "dilation" of the crystal lattice, or the strain. As the longitudinal wave propagates through the crystal, atomic planes are pushed closer to each other in some regions, and pushed farther apart in others, shown in Fig. 22.22.

Since smaller lattice constant semiconductors have larger bandgaps, the conduction band edge in compressively strained regions shift above their equilibrium values. Bardeen and Shockley (Fig. 22.23) modeled this in an intuitive manner by assuming that the conduction band edge energy shift is linearly proportional to the strain, giving a perturbation

$$W_\beta^{ADP}(r,t) = \Delta E_c(r) = D_a \nabla \cdot \mathbf{u}_\beta(\mathbf{r}) = iD_a \beta u_\beta e^{i(\vec{\beta}\cdot\mathbf{r}-\omega_\beta t)}, \tag{22.13}$$

where the proportionality constant D_a is called the conduction band **deformation potential**. A typical conduction band-edge deformation potential is of the order of $D_a \sim 10$ eV, which are typical band widths in the electronic bandstructure. The strain is $\sim u_\beta / a_0$ where a_0 is the lattice constant and Equation 22.9 gives u_β. Since strain is a **tensor**, so is the deformation potential. We restrict the discussion here to the simple isotropic cases, for which D_a is a scalar.

A valence band deformation potential is similarly defined. The *total* bandgap deformation potential is the sum of the conduction and valence band deformation potentials. It is measured in photoluminescence measurements as the shift of the semiconductor bandgap under externally applied strain. Conduction and valence band deformation potentials are indirectly accessible from transport measurements of n-type and p-type carriers respectively. It is straightforward to calculate the deformation potentials directly from the electronic bandstructure by evaluating the shift of band-edge energy eigenvalues by varying the lattice constant. This may be done most simply in the tight-binding bandstructure evaluations of Chapter 11, and is typically more frequently employed in the $\mathbf{k}\cdot\mathbf{p}$ bandstructure of Chapter 12.

We will now express the spatial part of every phonon scattering potential in the generalized form

$$W_+(r,t) = K_\beta u_\beta \cdot e^{+i\vec{\beta}\cdot\mathbf{r}} \cdot e^{-i\omega_\beta t} \text{ and } W_-(r,t) = K_\beta u_\beta \cdot e^{-i\vec{\beta}\cdot\mathbf{r}} \cdot e^{+i\omega_\beta t}, \tag{22.14}$$

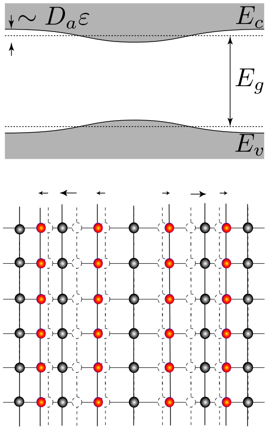


Fig. 22.22 Acoustic deformation potential perturbation to the band edges is caused by the compression and dilation of the lattice.

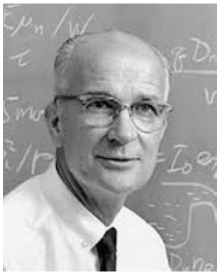


Fig. 22.23 William Shockley was a pioneer in semiconductor physics and devices. He formulated some of the earliest theories to explain electron transport phenomena in semiconductors, and with Bardeen formulated the theory of deformation potential scattering. He explained the curious phenomenon of the saturation of current in semiconductors by invoking phonon scattering. He was awarded the Nobel Prize in Physics in 1956 with Bardeen and Brattain for the invention of the transistor, the device that launched the information age.

---

[11] If all atoms are displaced equally, then the whole crystal has moved, and there is no compression or dilation. There is compression or dilation of the crystal if adjacent displacements are unequal, and thus the derivative $\partial u/\partial x$ is representative of such strain.

where + is for phonon absorption and − for phonon emission. Because $W_\pm$ has energy units, K_β has units of energy per length. For LA phonon ADP scattering, we read off K_β from Equation 22.13:

$$\boxed{K_\beta = iD_a\beta} \text{ for ADP scattering.} \tag{22.15}$$

For TA phonons, since $\mathbf{u}_\beta(\mathbf{r}) = \hat{\beta}_T u_\beta e^{i(\vec{\beta}\cdot\mathbf{r}-\omega_\beta t)}$, the direction of atomic vibrations $\hat{\beta}_T$ is *perpendicular* to $\vec{\beta}$, the direction of the wave propagation. Thus the strain is $\nabla\cdot\mathbf{u}_\beta(\mathbf{r}) = \hat{\beta}_T\cdot\vec{\beta}\times(...) = 0$, and thus there is no coupling with electrons, and no scattering[12].

Piezoelectric potential (PZ): When LA phonons propagate in a polar crystal, the strain field $\nabla\cdot\mathbf{u}_\beta(\mathbf{r})$ creates alternating planes of positive and negative charges because of the piezoelectric effect. The linear **piezoelectric coefficient** of polar semiconductors is defined so that the net dipole moment per unit volume, or the polarization magnitude is given by $P = e_{pz}\nabla\cdot\mathbf{u}_\beta(\mathbf{r})$. Though e_{pz} in general is a tensor, we discuss the scalar case[13]. The units of e_{pz} are in charge per unit area. Because the net charge in the semiconductor is zero, Maxwell's equations give, for the scalar case,

$$D = \epsilon_{dc}E + e_{pz}\frac{\partial u_\beta(r,t)}{\partial r} \text{ with } D = 0 \implies E(r,t) = -\frac{e_{pz}}{\epsilon_{dc}}\frac{\partial u_\beta(r,t)}{\partial r},$$
$$W_\beta^{PZ}(r,t) = -q\int drE(r,t) = \frac{qe_{pz}}{\epsilon_{dc}}u_\beta e^{i(\vec{\beta}\cdot\mathbf{r}-\omega t)} \implies \boxed{K_\beta = \frac{qe_{pz}}{\epsilon_{dc}}} \text{ for PZ.} \tag{22.16}$$

There is no piezoelectric scattering in non-polar semiconductors such as silicon and germanium. The TA modes do not couple to electrons through the piezoelectric component for the same reason there is no ADP coupling of transverse modes[14].

Optical deformation potential (ODP): Since the atomic planes vibrate out of phase in optical modes, the local strain is proportional, not to the differences in displacement captured by $\nabla\cdot\mathbf{u}_\beta$, but to the nearest neigbor atomic displacement $\mathbf{u}_s - \mathbf{u}_{s+1}$, which is the displacement vector $\mathbf{u}_\beta(\mathbf{r})$ itself[15]. The **optical deformation potential** D_{op} is thus defined as the proportionality constant

$$W_\beta^{ODP}(r,t) = D_{op}u_\beta e^{i(\vec{\beta}\cdot\mathbf{r}-\omega_\beta t)} \implies \boxed{K_\beta = D_{op}} \text{ for ODP scattering.} \tag{22.17}$$

A typical D_{op} is of the order of bandgap over lattice constant $\sim E_g/a_0$, which for most semiconductors is of the order of eV/Å, or 10^8 eV/cm.

Polar optical phonon (POP): In polar crystals, the optical modes produce effective charges $q^\star$ as discussed in Section 22.3. There we obtained the relation $\epsilon_{dc} - \epsilon_\infty = \frac{N(q^\star)^2}{\epsilon_0 M_r\omega_{TO}^2}$, with the mass density $\rho =$

[12] Fig. 22.7 shows this vividly!

[13] A dimensionless electromechanical coupling coefficient K is defined via the relation $\frac{K^2}{1-K^2} = \frac{e_{pz}^2/\epsilon_{dc}}{\rho v_s^2} = \frac{e_{pz}^2}{\epsilon_{dc}\rho v_s^2}$, which is the ratio of the electrical energy per unit volume e_{pz}^2/ϵ_{dc} to the mechanical energy per unit volume ρv_s^2. Typically $K^2 < 1$ for semiconductors. A higher K^2 is desired for the BAW and SAW devices that were discussed in Section 22.5.

[14] Material properties relevant to electron-phonon interactions are listed for various semiconductors in Tables A.2 and A.3 in the Appendix.

[15] ODP scattering of electrons for small k in the Γ-valley of direct bandgap semiconductors is forbidden by symmetry-constrained selection rules. It is allowed if the band minima are not at the Γ-point.

M_r/Ω where M_r is the reduced mass of atoms in a unit cell of volume Ω, and $N = 1/\Omega$. This yields the **Born effective charge**

$$q^\star = \Omega\omega_{TO}\sqrt{\rho(\epsilon_{dc} - \epsilon_\infty)} = \Omega\omega_{LO}\epsilon_\infty\sqrt{\rho(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})}. \tag{22.18}$$

The second form results from the first using the Lyddane–Sachs–Teller relation of Equation 22.8. The scattering potential is obtained in a manner similar to piezoelectric scattering. The microscopic charges have a dipole moment $p = q^\star u_\beta$, which produce a net volume polarization $P = q^\star u_\beta/\Omega$. Using this in Maxwell's equation yields

$$D = \epsilon_\infty E + \frac{q^\star u_\beta}{\Omega} \text{ with } D = 0 \implies E(r,t) = -\frac{q^\star u_\beta}{\epsilon_\infty \Omega}$$

$$\implies W_\beta^{POP}(r,t) = -q\int dr E(r,t) = \frac{q}{i\beta\epsilon_\infty}\cdot\frac{q^\star}{\Omega}\cdot u_\beta e^{i(\vec{\beta}\cdot\mathbf{r} - \omega_\beta t)}$$

$$\implies \boxed{K_\beta = \frac{q\cdot q^\star}{i\beta\epsilon_\infty} = \frac{q\omega_{LO}}{i\beta}\sqrt{\rho(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})}} \text{ for POP scattering.} \tag{22.19}$$

The polar optical phonon scattering potential is called the **Fröhlich potential** (Fig. 22.24). Unlike the deformation potentials, this potential is obtained entirely in terms of known macroscopic physical properties of the semiconductor crystal. The polar optical phonon scattering potential therefore has no unknown parameters and is accurate.

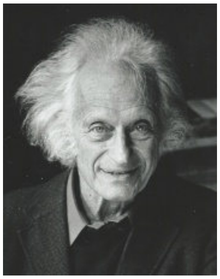


Fig. 22.24 Herbert Fröhlich performed seminal work on understanding how electrons interact with phonon modes in crystals. His work provided new insights to the quantum mechanics of transport phenomena in semiconductors as well as in understanding the origin of superconductivity.

Geometry of electron-phonon scattering: Since we know the perturbation potentials due to phonons that are experienced by electrons, we are in a position to compute the scattering rates. If a phonon in mode $\vec{\beta}$ scatters an electron from state $|\mathbf{k}\rangle \to |\mathbf{k}'\rangle$, momentum conservation requires $\mathbf{k}' = \mathbf{k} + \vec{\beta}$ for phonon absorption, and $\mathbf{k}' = \mathbf{k} - \vec{\beta}$ for phonon emission, neglecting Umklapp processes. Energy conservation requires $E_{\mathbf{k}'} = E_{\mathbf{k}} + \hbar\omega_\beta$ for phonon absorption and $E_{\mathbf{k}'} = E_{\mathbf{k}} - \hbar\omega_\beta$ for phonon emission. The electron experiences phonon scattering potentials $W_+(r,t) = K_\beta u_\beta e^{+i\vec{\beta}\cdot\mathbf{r}}e^{-i\omega_\beta t}$ for absorption and $W_-(r,t) = K_\beta u_\beta e^{-i\vec{\beta}\cdot\mathbf{r}}e^{+i\omega_\beta t}$ for emission. Fermi's golden rule immediately yields the absorption and emission rates as

$$\frac{1}{\tau_{\mathbf{k}}^{\text{abs}}(\beta,k')} = \frac{2\pi}{\hbar}|\langle\mathbf{k}'|K_\beta u_\beta e^{+i\vec{\beta}\cdot\mathbf{r}}|\mathbf{k}\rangle|^2\delta[E_{\mathbf{k}'} - (E_{\mathbf{k}} + \hbar\omega_\beta)], \text{ and}$$
$$\frac{1}{\tau_{\mathbf{k}}^{\text{em}}(\beta,k')} = \frac{2\pi}{\hbar}|\langle\mathbf{k}'|K_\beta u_\beta e^{-i\vec{\beta}\cdot\mathbf{r}}|\mathbf{k}\rangle|^2\delta[E_{\mathbf{k}'} - (E_{\mathbf{k}} - \hbar\omega_\beta)]. \tag{22.20}$$

Table 22.2 Phonon scattering matrix elements.

	K_β	$\lvert K_\beta\rvert^2\lvert u_\beta^{abs}\rvert^2$
ADP	$iD_a\beta$	$\frac{\beta^2 D_a^2\hbar N_\beta}{2\rho L^3\omega_\beta}$
PZ	$\frac{qe_{pz}}{\epsilon_{dc}}$	$\frac{q^2 e_{pz}^2\hbar N_\beta}{2\epsilon_{dc}^2\rho L^3\omega_\beta}$
ODP	D_{op}	$\frac{D_{op}^2\hbar N_\beta}{2\rho L^3\omega_\beta}$
POP	$\frac{qq^\star}{i\beta\Omega\epsilon_\infty}$	$\frac{q^2\omega_{LO}^2\hbar N_\beta}{2\beta^2 L^3\omega_\beta}(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})$

We postpone the complete evaluation of these scattering rates to Chapter 23, where the effect of phonons on electron mobility will be discussed. The scattering potential obtained for the four scattering modes: ADP, PZ, ODP, and POP are summarized in Table 22.2 and carried forward and expanded in Table 23.6 in Chapter 23. The squared matrix element involves $|K_\beta|^2|u_\beta|^2$. The quantum version of $|u_\beta|^2$ for absorption in Table 22.2 is obtained from Equation 22.9.

A few things can be noted about the dependence of the squared matrix elements $|K_\beta|^2|u_\beta|^2$ on the phonon wavevector β. For acoustic phonons of energy $\hbar\omega_\beta = \hbar v_s \beta << k_b T$, the Bose–Einstein occupation function $N_\beta \approx N_\beta + 1 \approx k_b T/\hbar v_s \beta >> 1$, resulting in a β^2 factor in the denominator. For ADP scattering, this cancels the β^2 dependence in the numerator, making the squared matrix element independent of the phonon wavelength. But for PZ scattering, the matrix element is much larger for longer wavelength phonons as it goes as $1/\beta^2$. For optical phonons, it may be assumed that $\omega_\beta \approx \omega_0$, in which case N_β is independent of β. Then, the squared matrix element is independent of β for ODP scattering, but for POP scattering it goes as $1/\beta^2$, similar to PZ scattering. This is expected because unlike ADP and ODP scattering which originate from deformation, both PZ and POP scattering originate from charges whose Coulomb forces have a long range. Long-range forces typically strongly favor scattering of electrons by small angles, as we describe next.

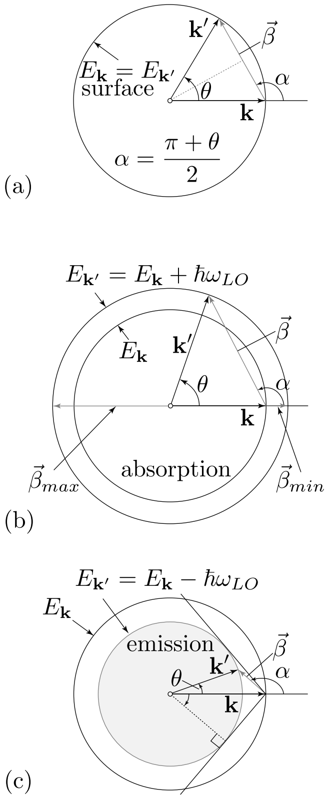


Fig. 22.25 Phonon scattering geometry for (a) elastic phonon emission and absorption, (b) inelastic phonon absorption, and (c) inelastic phonon emission.

Table 22.3 Allowed phonon β ranges.

	Acoustic Absorption	Emission
$\frac{\beta_{max}}{k}$	0	$2(1 - \frac{m^\star v_s}{\hbar k})$
$\frac{\beta_{min}}{k}$	$2(1 + \frac{m^\star v_s}{\hbar k})$	0
	Optical Absorption	Emission
$\frac{\beta_{max}}{k}$	$\sqrt{1 + \frac{\hbar\omega_0}{E_\mathbf{k}}} + 1$	$1 + \sqrt{1 - \frac{\hbar\omega_0}{E_\mathbf{k}}}$
$\frac{\beta_{min}}{k}$	$\sqrt{1 + \frac{\hbar\omega_0}{E_\mathbf{k}}} - 1$	$1 - \sqrt{1 - \frac{\hbar\omega_0}{E_\mathbf{k}}}$

For the complete evaluation of the scattering rates and mobility in Chapter 23, the last ingredient required is an accurate description of the geometry enforced in the electron-phonon scattering events to satisfy momentum and energy conservation relations, which is discussed here. Fig. 22.25 shows a phonon in mode $\vec{\beta}$ scattering an electron from state $|\mathbf{k}\rangle$ into a state $|\mathbf{k}'\rangle$ for (a) elastic scattering, which is valid for low-energy acoustic phonons, and (b, c) for inelastic scattering, which is valid for optical phonons. As seen before, momentum conservation requires $\mathbf{k}' = \mathbf{k} + \vec{\beta}$ for phonon absorption, and $\mathbf{k}' = \mathbf{k} - \vec{\beta}$ for phonon emission, neglecting Umklapp processes. Energy conservation requires $E_{\mathbf{k}'} = E_\mathbf{k} + \hbar\omega_\beta$ for phonon absorption and $E_{\mathbf{k}'} = E_\mathbf{k} - \hbar\omega_\beta$ for phonon emission.

Consider first the geometry of elastic scattering shown in Fig. 22.25 (a). The initial and final electron wavevectors $(\mathbf{k}, \mathbf{k}')$ lie on a constant energy sphere for 3D transport (angles for lower dimensions are obtained in a similar fashion). The angle between the phonon wavevector $\vec{\beta}$ and the initial electron wavevector $\mathbf{k}$ is α, which is related to θ, the angle by which the electron is scattered, by $\alpha = (\pi + \theta)/2$. So when $\theta : 0 \to \pi$, $\alpha : \frac{\pi}{2} \to \pi$. It is clear that if the acoustic phonon energy is neglected compared to the electron energy, the magnitude of the phonon wavevectors are between $0 \leq \beta \leq 2k$. This is an excellent approximation for acoustic phonons. We will find the exact range shortly, but we can say more about the nature of scattering before doing that. Since the squared matrix element for ADP scattering in Table 22.2 is independent of β, all $0 \leq \beta \leq 2k$ have the same scattering strength. Thus the electron is scattered over all angles $0 \leq \theta \leq \pi$ with equal preference: this is **isotropic** scattering. On the other hand, for acoustic PZ phonon scattering, the squared matrix element goes as $1/\beta^2$, which is

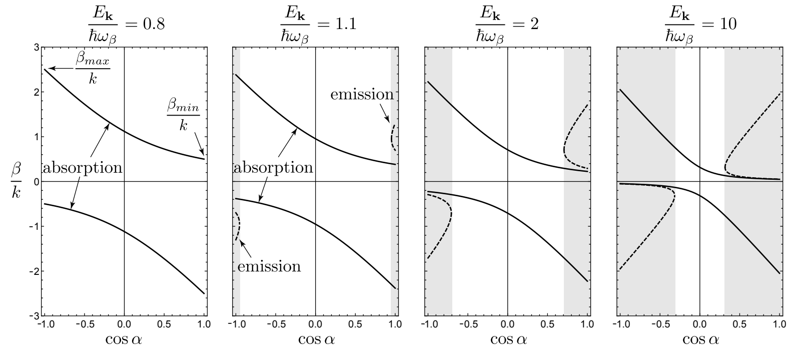


Fig. 22.27 A plot of Equation 22.23 highlighting constraints on emission and absorption of phonons by electrons. The panels show successively increasing electron kinetic energy E_k compared to the phonon energy $\hbar\omega_\beta$. For $E_k < \hbar\omega_\beta$, no emission is allowed.

large for small β, implying that small α and small θ are heavily favored and the scattering is not isotropic.

We use the convention that upper signs are for absorption and lower signs are for emission. For a parabolic electron energy bandstructure $E_k = (\hbar k)^2/(2m^\star)$, equating the momentum and energy before and after scattering with the angles of Fig. 22.25 (a, b, c), we get

$$\mathbf{k}' = \mathbf{k} \pm \vec{\beta} \implies (k')^2 = k^2 + \beta^2 \pm 2\beta k \cos\alpha$$
$$E_{\mathbf{k}'} = E_{\mathbf{k}} \pm \hbar\omega_\beta \implies (\frac{\beta}{k})^2 \pm 2\frac{\beta}{k}\cos\alpha \mp \frac{\hbar\omega_\beta}{E_k} = 0. \tag{22.21}$$

For acoustic phonons, $\hbar\omega_\beta = \hbar v_s \beta$, and we get

$$\frac{\beta}{k} = 2(\mp\cos\alpha \pm \frac{m^\star v_s}{\hbar k}) = 2(\mp\cos\alpha \pm \frac{v_s}{v_k}), \tag{22.22}$$

where $v_k = \hbar k/m^\star$ is the electron group velocity.

For a generic optical phonon dispersion ω_β, we get

$$\frac{\beta}{k} = \mp\cos\alpha \pm \sqrt{\cos^2\alpha \pm \frac{\hbar\omega_\beta}{E_k}}. \tag{22.23}$$

Fig. 22.27 shows that while absorption is allowed with all angles α, an electron may emit a phonon only if it has an energy E_k more than the optical phonon energy $\hbar\omega_\beta$. This restricts the allowed angles. As the kinetic energy E_k of the electron increases well beyond the optical phonon energy, the scattering event starts appearing more like the elastic version where all angles are allowed[16].

Equations 22.22 and 22.23 with the bounds $-1 \leq \cos\alpha \leq +1$ and the restriction of real β set the exact ranges of the magnitude of β/k allowed in phonon scattering. Consider the geometry of electron scattering from longitudinal optical (LO) phonons. A characteristic wavevector k_{LO} is defined via the relation $(\hbar^2 k_{LO}^2)/(2m^\star) = \hbar\omega_{LO}$. **For phonon absorption** shown in Fig. 22.25 (b), the electron energy difference leads to $(k')^2 = k^2 + k_{LO}^2$. The angle α can take all values between $(0, \pi)$, but

Fig. 22.26 Cherenkov observed that electrons moving though dielectric media faster than the velocity of light *in that medium* emitted a radiation, or glow – which is named after him, and for which he was awarded the 1958 Nobel Prize in Physics. The acoustic analogs are the boom of a supersonic jet, and the emission of acoustic phonons by electrons.

[16] From Fig. 22.21 (b), there is a minimum energy required for electrons to emit acoustic phonons too. Though acoustic phonon scattering is considered nearly elastic, an electron cannot emit acoustic phonons unless it is moving faster than the speed of sound v_s in the semiconductor, which is of the order of 10^5 cm/s. Such energy and angle restrictions have direct analogy to Cherenkov radiation, which occurs when electrons move through media faster than the phase velocity of light (see Fig. 22.26).

the phonon wavevectors range in magnitude from $\beta_{min}^{abs} = k' - k = k[\sqrt{1 + (k_{LO}^2/k^2)} - 1]$ to $\beta_{max}^{abs} = k' + k = k[\sqrt{1 + (k_{LO}^2/k^2)} + 1]$, as can be read off from the figure. **For phonon emission** shown in Fig. 22.25 (c), $(k')^2 = k^2 - k_{LO}^2$, which leads to $\beta_{min}^{em} = k - k' = k[1 - \sqrt{1 - (k_{LO}^2/k^2)}]$ and $\beta_{max}^{em} = k + k' = k[1 + \sqrt{1 - (k_{LO}^2/k^2)}]$. Clearly $E_{\mathbf{k}} \geq \hbar\omega_{LO}$, or equivalently $k \geq k_{LO}$ for an electron to be able to emit an optical phonon. The tangent from the tip of the vector $\mathbf{k}$ to the smaller circle defines a maximum angle θ, which implies that α ranges from π to a maximum value given by $\cos\alpha_{max} = k_{LO}/k$. Table 22.3 shows the resulting ranges for β for both acoustic and optical phonon scattering. We will return to these considerations in Chapter 23, where we will evaluate phonon scattering limited electron mobility.

22.9 Chapter summary section

In this chapter, we learned that:

- The effect of phonons in crystals on the transport properties of light and thermal properties of heat, were instrumental in the development of quantum mechanics.
- Phonon modes are lattice vibrations that have energies less than $10 - 100$ meV for most semiconductors. Phonons are bosons. Their occupation function follows the Bose–Einstein distribution function.
- Adjacent atomic planes can vibrate in-phase, or out of phase, leading to acoustic and optical phonons respectively. For N atoms in the primitive cell, are $3N$ phonon branches of which 3 are acoustic branches, and $3N - 3$ are optical modes.
- Acoustic phonons have a zero energy mode with the dispersion roughly $\hbar\omega_\beta = \hbar v_s \beta$ where v_s is the sound velocity. Optical modes have small energy dispersion and are sometimes approximated to be dispersionless ($\omega_\beta \approx \omega_0$).
- Photons interact strongly with the TO phonon modes. The optical reflectivity at frequencies between the TO and LO energies is nearly unity. This is knowns as the Reststrahlen band, and is a direct measurement of the two energies.
- Photons can absorb or emit quanta of phonons upon reflection from a semiconductor. This inelastic scattering of photons by phonons is called Raman scattering for optical modes, and Brillouin scattering for acoustic modes. Both techniques are used extensively to identify (or ”fingerprint”) semiconductors, and to also probe their strain or stress states.
- Inelastic neutron scattering, or X-ray scattering is used to measure the entire phonon dispersion experimentally.

- Acoustic wave propagation is the principle behind SAW and BAW filters used in RF and microwave electronics.
- The thermal conductivity of semiconductors is controlled primarily by phonons. Light atoms semiconductors with few atoms in the unit cell have high thermal conductivities.
- Electrons scatter quasi-elastically from acoustic phonons, and inelastically from optical phonons. The scattering rates are evaluated using Fermi's golden rule applied individually to various phonon modes.
- The four ways phonons scatter electrons are via the acoustic deformation potential (ADP), the acoustic piezoelectric interaction (PZ), the optical deformation potential (ODP), and the polar optical phonon (POP).

Further reading

The classic reference for phonons in crystals is **Dynamical Theory of Crystal Lattices** by M. Born and K. Huang. The series **Light Scattering by Solids** edited by M. Cardona covers the many facets of interaction of photons with phonons. Ziman's **Electrons and Phonons** has an unparalleled treatment of electron-phonon interactions looked at from various angles and has Ziman's signature writing style that is quite a joy to read. **Carrier Scattering in Metals and Semiconductors** by Gantmakher and Levinson has a comprehensive description of electron-phonon interactions in metals and semiconductors. **Electrons and Phonons in Semiconductors Multilayers** by B. Ridley and **Phonons in Nanostructures** by M. Stroscio and M. Dutta are texts that go deeper into electron-phonon interactions in semiconductors and their heterostructures.

Exercises

(22.1) **Phonons carry no net momentum**

Just as the vibrations of water molecules in a closed bottle do not move the bottle, phonons are internal vibrations of the atoms of a crystal, they do not carry net momentum and do not move the entire crystal. To prove this, note that the NET momentum carried by all phonons in a crystal is

$$M\frac{d}{dt}\sum_s u_s. \qquad (22.24)$$

Since the atomic displacements are of a plane-wave nature $u_s \propto e^{iKsa}$, show that the net momentum is

$$p = M\frac{du}{dt}\sum_s e^{isKa} = M\frac{du}{dt}, \frac{1-e^{iNKa}}{1-e^{iKa}}, \qquad (22.25)$$

and since for phonon wavevectors $K = \pm 2\pi r/Na$, show that the total momentum carried by the crystal due to phonons is *zero*. For $K = 0$, $p = NM\frac{du}{dt}$ represents the case when the whole crystal (all the N atoms) are moved at a velocity $\frac{du}{dt}$.

(22.2) **Raman spectra and the Lyddane–Sachs–Teller re-**

lation for GaN and AlN
Fig. 22.17 showed the Raman spectra measured for GaN and AlN, two technologically important wide bandgap semiconductor materials.

(a) Using the phonon dispersion of GaN shown in Fig. 22.10, identify the phonon modes responsible for each of the three peaks in the Raman spectra.

(b) Identify ω_{TO} and ω_{LO} for GaN. If it is known that the DC relative dielectric constant of GaN is $\epsilon_{dc} = 8.9$, use the Lyddane–Sachs–Teller relation to find the high-frequency dielectric constant ϵ_∞. Confirm by searching the literature if the value you calculate is indeed experimentally observed.

(c) Assigning the same ordering to the optical phonon peaks observed in the Raman spectra of AlN, repeat the above procedure for the dielectric constants of AlN.

(d) Explain why the phonon frequencies of AlN are larger than the corresponding frequencies in GaN. Confirm that the toy model of Fig. 22.9 is reasonably accurate and can be used to predict this trend at least qualitatively.

(22.3) **Acoustic wave filters**
Discuss how surface acoustic wave (SAW) and bulk acoustic wave (BAW) filters are used in high-frequency communication systems. SAW filters operate up to ~ 1 GHz and BAW filters from $1-10$ GHz. Calculate and plot the center frequency vs. the cavity length for BAW filters for various sound velocities. In addition to the sound velocity an important metric is the electro-mechanical coupling coefficient. Make a table of the materials that allow for the highest performance BAW filters.

(22.4) **Calculating the thermal conductivity from the Boltzmann transport equation**
The Boltzmann transport equation of Chapter 21 provides the recipe to evaluate the thermal conductivity of solids in analogy to the electrical conductivity.

(a) Denoting heat current as Q, the nonequilibrium phonon number n_β of phonon mode β and equilibrium phonon number n_0, the phonon energy of $\hbar\omega_\beta$ and group velocity v_β, show that the heat current is

$$Q = \frac{1}{L^d}\sum_\beta (n_\beta - n_0)\hbar\omega_\beta v_\beta = -\underbrace{\left(\frac{1}{L^d}\sum_\beta \frac{\partial n_{ph}}{\partial T}\hbar\omega_\beta v_\beta\right)}_{\kappa}\frac{\partial T}{\partial x}, \tag{22.26}$$

which is the Fourier law for heat current with thermal conductivity κ.

(b) Using further a phonon relaxation time of τ_{ph} show that the thermal conductivity is obtained in the steady state when $\frac{\partial}{\partial t}(...) = 0$ as

$$\frac{dn_{ph}}{dt} = \underbrace{v_\beta \frac{\partial n_{ph}}{\partial T}\frac{\partial T}{\partial x}}_{\text{inflow}} - \underbrace{\frac{n_{ph}-n_0}{\tau_{ph}}}_{\text{decay}}$$

$$\implies n_{ph} = n_0 + v_\beta \tau_{ph}\frac{\partial n_{ph}}{\partial T}\frac{\partial T}{\partial x}$$

$$\implies \boxed{\kappa = \frac{1}{3}c_v v\lambda = \frac{1}{3}c_v v^2\tau.} \tag{22.27}$$

Note that this is identical to the classical result that was discussed for the thermal conductivity due to conduction electrons in metals in Chapter 2, Equation 2.7. But the velocity v, specific heat c_v and scattering time τ here are those of phonons, not of free–electrons.

(22.5) **Thermal boundary resistance**
Like light and electrons, phonons undergo reflections and transmission at interfaces. Discuss how this results in a **thermal boundary resistance** (TBR). Tabulate and compare the values of the TBR for several important interfaces, such as Si/SiO$_2$ and GaN/SiC. Discuss how the TBR is a major bottleneck for several high-performance electronic and photonic devices.

(22.6) **Normal is not enough: need for Umklapp**
Consider an electron distribution that is carrying current in a semiconductor in response to an external electric field. If at time $t = 0$ the field is removed, the electron distribution must return to its equilibrium Fermi–Dirac distribution state. However, it is impossible for this to occur with only normal electron-phonon scattering events; Umklapp processes are necessary for return to equilibrium. A brief proof is outlined here:

(a) A normal electron-phonon scattering process follows the momentum conservation relation

$\mathbf{k}_i \pm \mathbf{q} = \mathbf{k}_f$, where $\hbar\mathbf{k}_i$ is the initial electron momentum, $\hbar\mathbf{k}_f$ is the final electron momentum and $\hbar\mathbf{q}$ is the phonon momentum. The energy conservation relation accompanying this process is $E_{el}(\mathbf{k}_f) = E_{el}(\mathbf{k}_i) \pm E_{ph}(\mathbf{q})$ where the electron and phonon energies are written out explicitly. Since an Umklapp process involves a phonon of wavevector $\mathbf{G}$ where $\mathbf{G}$ is a reciprocal lattice vector, the momentum conservation relation changes to $\mathbf{k}_i \pm \mathbf{q} + \mathbf{G} = \mathbf{k}_f$. How is the energy conservation relation modified?

(b) Discuss, based on the Boltzmann transport equation, why Umklapp processes are needed to bring a non-equilibrium distribution of electrons back to equilibrium.

(c) At very low temperatures Umklapp scattering of electrons in metals is the dominant process leading to electrical resistivity. Discuss why this is the case if the Fermi energy E_F spans a significant fraction of the Brillouin zone.

(d) Discuss why Umklapp processes are responsible for the high-temperature thermal conductivity of sufficiently pure electrically insulating crystals.

(22.7) **Quantum of thermal conductance**

In Chapter 17 we saw the quantum limit of the maximum electrical conduction across a single electronic channel is the electrical conductance quantum $G_{1d} = q^2/h$, where q is the unit charge and h is the Planck constant, which is confirmed experimentally. There is also a similar quantum limits on the thermal conductance. In this problem, you calculate the *thermal* conductance quantum, and show that the Wiedemann–Franz law of the classical Drude model also holds in the quantum limit.

For a 1D conductor connected to a source contact on the left, and a drain contact on the right which is at a voltage V, the right flowing current is $J^R_{1d} = q\frac{g_s g_v}{L}\sum_{k\geq 0} v_g(k) f_s(k)$ where all symbols have their usual meanings. Now suppose that in addition to the voltage difference, there is also a temperature gradient across the conductor i.e. the source is at a slightly higher temperature $(T_0 + \Delta T)$ than the drain (T_0). The temperature gradient gives rise to a *heat* current. The heat carried by the electron in state $\mathbf{k}$ is given by $(E(k) - \mu)$ where μ is the chemical potential of that electronic state. Analogous to the charge current, the right flowing heat current is $Q^R_{1d} = q\frac{g_s g_v}{L}\sum_{k\geq 0}(E(k) - E_{Fs}) v_g(k) f_s(k)$.

(a) Calculate the exact net heat current $Q_{1d} = Q^R_{1d} - Q^L_{1d}$ in the 1D ballistic device.

(b) Take the degenerate limit of the net heat current assuming a high carrier density and $g_s = g_v = 1$.

(c) Now the electrical bias between the source and drain is removed. Show that the thermal conductance $G_Q = Q_{1d}/\Delta T = \pi^2 k_b^2 T_0/3h$ in 1D for a small temperature difference ($\Delta T \ll T_0$) in the degenerate limit. Calculate the numerical value of G_Q at $T_0 = 4$ K and 300 K with proper units. This thermal conductance quantum sets the maximum thermal conduction across a single transport channel.

(d) Calculate the ratio $G_Q/G_{1d}T_0$, and show that you obtain the Lorenz number. This is the quantum form of the Wiedemann–Franz law in the ballistic limit. Compare it with the Wiedemann–Franz law in the classical Drude picture.

We have now seen that electrons produce both charge and heat currents. In metals electrons carry most of the heat, but in semiconductors, electrons carry only a part of the heat; most of the heat is carried by lattice vibrations, or phonons. The acoustic phonon velocity (or sound velocity) v_s is the slope of the acoustic phonon energy dispersion $E(k) = \hbar v_s k$. The acoustic modes transport most of the heat. Taking the same 1D semiconductor ballistic device as in this problem, now calculate the heat carried by long wavelength acoustic phonons. Phonons obey Bose–Einstein statistics and their chemical potential is 0.

(e) Calculate the net heat current $Q_{1d,ph}$ carried by phonons as you did for electrons, and calculate the thermal conductance G_Q in 1D for small temperature difference ($\Delta T \ll T_0$), and find $G_{Q,ph} = Q_{1d,ph}/\Delta T$. You should again obtain the thermal conductance quantum $G_{Q,ph} = \pi^2 k_b^2 T_0/3h$.

(f) This behavior was experimentally confirmed in 2000. Read the paper [K. Schwab, E. Henriksen, J. Worlock, and M. L. Roukes, Measurement of the quantum of thermal conductance, Nature 404, 974 (2000)], and comment on what is measured and what you have calculated.

Note: From the above it is found that rather surprisingly, $G_Q = G_{Q,ph}$. Fermions and bosons follow different statistics, yet electrons and phonons end up giving the same thermal conductance quantum in 1D. [This problem was created by Jashan Singhal of Cornell University. See Physical Review Research, volume 2, page 043413 (2020).]

(22.8) **Simplified explanation of phonon scattering**
In this problem, a heuristic derivation is given why the linear increase in electrical resistivity in metallic conduction in pure crystals is limited by phonon scattering.

(a) Because of thermal vibrations, an atom of mass M can be thought of as connected to springs. Because of the thermal energy $k_b T$, its location is blurred out over an oscillation amplitude u_0. Using $\frac{1}{2}M\omega^2 u_0^2 \approx k_b T$, the atom is blurred over a circle of area $\sigma = \pi u_0^2$. Estimate this area for say a silicon atom at room temperature.

(b) Show then that the mean free path of an electron before it collides with the thermally smeared out "atom-cloud" is $l_{mfp} \approx \frac{1}{\sigma} \propto \frac{1}{T}$.

(c) Since the electrical conductivity is $\sigma \propto l_{mfp}$, argue why $\sigma \sim \frac{1}{T}$ and $\rho \sim T$.

(22.9) **Strong electron-phonon coupling**
In this chapter, we identified the scattering potentials in Section 22.8. In the next Chapter (23) the scattering rates for 3D and 2D electrons by acoustic and optical phonons will be discussed in detail. This problem discusses the case of strong coupling of electrons and phonons that lead to superconductivity.

(a) Explore how strong coupling of electrons with phonons can make the interaction between two electrons switch from the standard Coulomb repulsion to – rather astonishingly – to an attractive interaction!

(b) The Bardeen–Cooper–Schrieffer (BCS) theory of superconductivity shows that the formation of a large number of composite electron-phonon-electron objects, called a **Cooper pair** is energetically more favorable at a low temperature than the electrons remaining independent. This is somewhat like the condensation of water from vapor into a liquid. Discuss why the electrical resistivity of such a superfluid vanishes. Discuss why this is somewhat counterintuitive: the same phonons that increase the resistivity in normal metallic transport, end up helping to lead to the complete loss of resistivity.

Scattering, Mobility, and Velocity Saturation

23

23.1 **Electron mobility: a résumé** 555
23.2 **Scattering mechanisms** 557
23.3 **Point defect scattering** 574
23.4 **Coulomb impurity scattering** 577
23.5 **Dipole scattering** 580
23.6 **Dislocation scattering** 581
23.7 **Alloy disorder scattering** 583
23.8 **Interface scattering** 586
23.9 **Phonon scattering** 588
23.10 **Experimental mobilities** 602
23.11 **High-field velocity saturation** 604
23.12 **Chapter summary section** 607
Further reading 607
Exercises 608

The mobility of electrons and holes plays a defining role in determining the suitability of a semiconductor for practical applications. Since the electrical conductivity $\sigma = qn\mu$ is the product of the mobile carrier density n and mobility μ, for most applications it is desired to have the highest possible mobility, to achieve the highest electrical conductivity with the least number of mobile electrons (or holes). In this chapter,

- We survey the experimental landscape of electron and hole mobilities in many scenarios ranging from zero/narrow bandgap semiconductors to ultra wide bandgap semiconductors,
- Investigate quantitatively the scattering mechanisms that limit the electron mobility, ranging from intrinsic processes such as phonons, to extrinsic processes such as dopants, alloy disorder, defects, and interfaces, and
- Investigate the drift velocity as a function of the external electric field to understand velocity saturation and high-field effects.

23.1 Electron mobility: a résumé

Fig. 23.1 (a) shows a typical trend of electron (or hole) mobilities measured in a semiconductor as a function of temperature, and (b) the mobility as a function of the doping density in the semiconductor. Though the exact numbers may vary, the general trends seen in this figure (with minor variations) are seen in most bulk doped semiconductors, and also in semiconductor heterostructures that house lower-dimensional conduction channels. Thus, the purpose of this chapter is to first qualitatively understand the reason behind these trends, and then quantitatively calculate the mobilities from the Boltzmann transport equation using Fermi's golden rule for the scattering rates.

The mobility at the lowest (cryogenic) temperatures is limited by defect and impurity scattering because both acoustic and optical phonons are frozen out, as discussed in Chapter 22. Lower impurity and defect densities lead to higher mobilities: for example in ultraclean GaAs 2DEGs, mobilities as high as $\sim 10^7$ cm^2/V·s have been achieved, resulting in mean free paths of several mm! At room temperature, the mobilities limited by phonon scattering fall in the $\sim 1000 - 40,000$

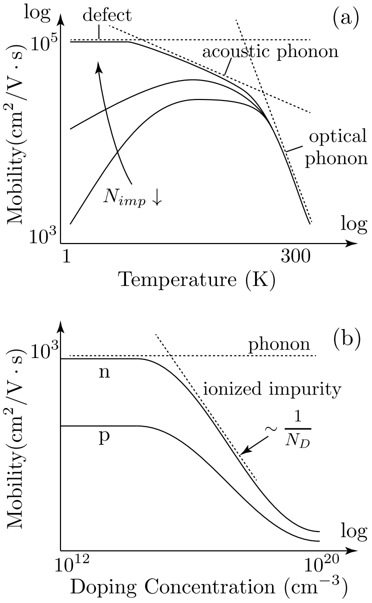


Fig. 23.1 Semiconductor mobility trends.

cm^2/V·s range for many useful semiconductors. In most lightly doped compound semiconductors, polar-optical phonon scattering is the dominant mechanism limiting the mobility at room temperature. In elemental non-polar semiconductors such as silicon and germanium this mode is absent, and acoustic phonons dominate. In 2D electron (or hole) gases confined at heterointerfaces such as the silicon MOSFET, the mobility is limited by scattering from the rough interface between the crystalline and amorphous material, limiting it to ~ 200 cm^2/V·s. But for similar 2DEGs in III-V epitaxial quantum wells, a crystalline heterojunction between nearly lattice-matched semiconductors allows the mobility to reach the intrinsic optical phonon limits, ranging from ~ 2000 cm^2/V·s in wide bandgap Al(Ga)N/GaN heterostructures to $\sim 40,000$ cm^2/V·s in narrow bandgap InSb quantum wells at 300 K.

Fig. 23.1 (b) shows that at room temperature, the mobility in a very lightly doped semiconductor will approach the phonon limit. But as the doping N_D is increased beyond a certain density, the Coulomb scattering from the ionized donors takes over. The ionized impurity scattering limited mobility decreases as $\sim 1/N_D$ within a window, before leveling off at the highest doping densities. In semiconductor devices, the heavily doped regions typically form ohmic contacts and access regions where a high conductivity is desired. The much higher carrier densities in such heavily doped regions offset the lower mobilities, and deliver the desired higher net conductivity.

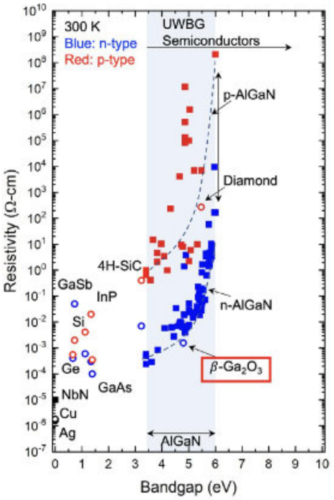


Fig. 23.2 Conductivity vs. bandgaps of various n-type and p-type doped semiconductors and comparison to metals. Kindly shared by Dr. Kevin Lee of Cornell University.

Fig. 23.2 shows the n-type and p-type resistivities reported for several semiconductors and metals vs. their energy bandgaps. The resistivities of some metals are also shown for comparison. Typical n-type resistivities are lower than p-type resistivities. The lowest semiconductor resistivities are about 200× higher than the lowest metallic resistivities. The general trend of increasing resistivity with bandgaps is only partially due to low mobilities: the more important challenge is to obtain enough mobile carriers due to deep donors or acceptors.

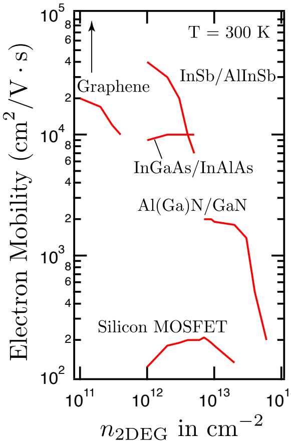


Fig. 23.3 2DEG mobilities vs. carrier densities at room temperature for various semiconductors.

Fig. 23.3 shows the room-temperature electron mobilities in 2D electron gases in some technologically important semiconductors that are used for field-effect transistors as a function of the 2D sheet density. The highest mobilities of electron mobilities of $\sim 10,000 - 40,000$ cm^2/V·s are achieved in quantum well heterostructures of narrow-bandgap semiconductors such as InGaAs/InAlAs and InSb/AlInSb. These are used in transistors which can also be turned on and off due to the presence of a bandgap, albeit narrow. In even narrower bandgap HgCdTe or zero-gap graphene, even higher mobilities can be achieved. But because of the negligible or zero bandgaps, it is not possible to modulate their conductivities enough for transistor action in the field-effect geometry. On the other hand typical 2DEG mobilities in silicon MOSFETs are in the ~ 200 cm^2/V·s range, and that in AlGaN/GaN 2DEGs in the 1500 cm^2/V·s range. In these semiconductors, the density of 2DEGs can be typically higher than the narrow-bandgap semiconductors because they can sustain higher electric fields without breaking down (see Fig. 23.4).

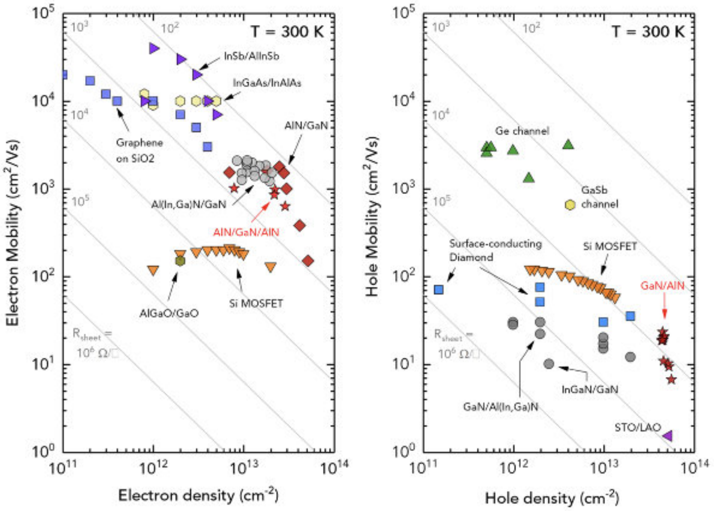


Fig. 23.4 Room temperature mobilities of 2D electron gases and 2D hole gases vs. the respective sheet carrier densities in various semiconductor families. These transport properties are of high importance in field-effect transistors (FETs) for several applications. Contours of constant sheet resistance are shown in the background, indicating the highest conductivities to the top right. Kindly shared by Dr. Reet Chaudhuri, Cornell University.

Fig. 23.5 shows typical electron and hole mobilities for various semiconductors as a function of the thickness of the semiconductor channel layer. In this case, it is seen that when the thickness approaches nanoscale dimensions, the mobility degrades because of roughness scattering. The class of 2D layered materials do not suffer from this roughness, and can achieve higher mobilities than conventional semiconductors of comparable layer thickness in the thin, nanoscale limit.

The range of transport phenomena exhibited by semiconductors is far richer than this short résumé. In this chapter we will be able to sample a large portion, but not all of them. Amorphous and organic semiconductors do not receive the same in-depth treatment as the inorganic semiconductors do in this book. Some semiconductors lose their resistivities completely, and undergo an electronic phase transition into superconductivity: this phenomena is not discussed either. In the next section, we describe the transport theory formalism that will be used to quantitatively calculate the experimental mobilities for various semiconductors. The mobilities will be found to depend on their intrinsic physical properties, such as the bandgaps, effective masses, dielectric constants, deformation potentials, etc. The mobility also depends on the physical conditions such as the temperature, and the defect and impurity densities. Based on these considerations, we classify the scattering mechanisms, before proceeding to evaluating individual scattering mechanisms, and combining them at the end to explain the experimental data described in this short résumé.

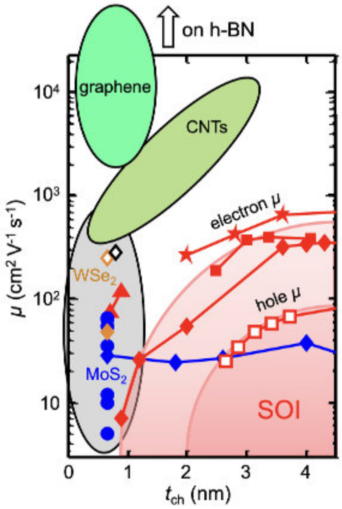


Fig. 23.5 Mobilities vs. channel thickness for various semiconductors. Filled red symbols are silicon electron mobility and empty symbols are silicon hole mobility. SOI stands for silicon on insulator, CNT stands for carbon nanotubes, and on h-BN means graphene on hexagonal boron nitride. For CNTs t_{ch} roughly corresponds to the CNT diameter as opposed to the layer thickness elsewhere. Modified figure from C. English, E. Pop et al. Nano Lett. 16, 3824 (2016). Kindly shared by Dr. Eric Pop, Stanford University.

23.2 Scattering mechanisms

The flowchart in Fig. 23.7 shows the method by which the mobility in semiconductors is evaluated using the relaxation-time approximation of the Boltzmann transport equation. Most steps that go into this evaluation have been discussed in earlier chapters. Here we summarize the steps succinctly and proceed directly to evaluate individual scattering rates and the resulting mobility. Some concepts that have not been discussed before, such as screening of scattering potentials are covered and illustrated in the evaluation of the mobility, while others are recalled and refined for the transport problem.

0. Carrier statistics: Because the conductivity $\sigma = qn\mu$ defines the mobility $\mu = \sigma/qn$, under no circumstance must one naively think that the mobility is *independent* of the carrier concentration n. The mobility depends on the carrier density n. Neglecting this basic fact sometimes leads to confusion and erroneous conclusions. For example, the mobility of electrons in nearly intrinsic GaAs at 300 K may reach 8000 $\text{cm}^2/\text{V}\cdot\text{s}$ when n is extremely low. If n is increased to $10^{20}/\text{cm}^3$ by donor doping, the mobility plummets to below 1000 $\text{cm}^2/\text{V}\cdot\text{s}$ due to ionized impurity scattering. On the other hand, for similar volume carrier concentrations realized in 2DEG sheets at AlGaAs/GaAs heterojunctions by field effect or by modulation doping, close to the intrinsic mobility of 8000 $\text{cm}^2/\text{V}\cdot\text{s}$ is achieved by the removal of ionized impurity scattering. Mobility is *not* a fundamental material parameter, and should be understood to be related to the carrier density in *every* context.

Furthermore, it is important to bear in mind that in semiconductor devices, the mobility typically is *fixed* once the semiconductor structure has been created, and the knob that makes the device click is not so much the mobility, but the carrier density n, which is tuned by several orders of magnitude either by field-effect in FETs, or by carrier injection in pn diodes, bipolar transistors, LEDs, or lasers. Because one needs mobile carriers in the first place to speak about their mobility, it is worthwhile recalling some concepts of carrier densities in bulk semiconductors, and quantized heterostructures such as quantum wells.

The density of molecules in air at room temperature and pressure is roughly $10^{19}/\text{cm}^3$. Since the density of electrons in semiconductors reaches similar values at the higher side, it is partly the reason it is referred to as an "electron gas". The concentration of mobile electrons in the conduction band depends on the energy separation of the Fermi level and the conduction band edge $\eta = \frac{E_F - E_c}{k_b T}$, which is made non-dimensional by dividing by the temperature. The carrier density in a parabolic band semiconductor where the electrons are free to move in d-dimensions is given by the following forms from earlier chapters:

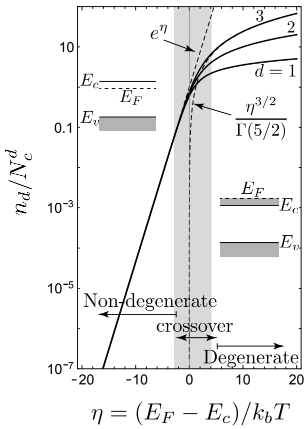


Fig. 23.6 The dependence of mobile electron density in d-dimensions on both the Fermi level E_F and on temperature via $\eta = \frac{E_F - E_c}{k_b T}$. A similar relation holds for mobile holes.

$$n_d = \frac{g_s g_v}{L^d} \sum_{\mathbf{k}} f^0_{\mathbf{k}} = g_s g_v \int \frac{d^d k}{(2\pi)^d} f^0_{\mathbf{k}} = \int_{E_c}^{\infty} dE_{\mathbf{k}} g_d(E_{\mathbf{k}}) f^0_{\mathbf{k}}$$

$$\underbrace{\Longrightarrow}_{E_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m^\star}} n_d = \underbrace{g_s g_v \left(\frac{2\pi m^\star k_b T}{h^2}\right)^{\frac{d}{2}}}_{N_c^d} F_{\frac{d}{2}-1}\left(\frac{E_F - E_c}{k_b T}\right) = N_c^d F_{\frac{d}{2}-1}(\eta), \quad (23.1)$$

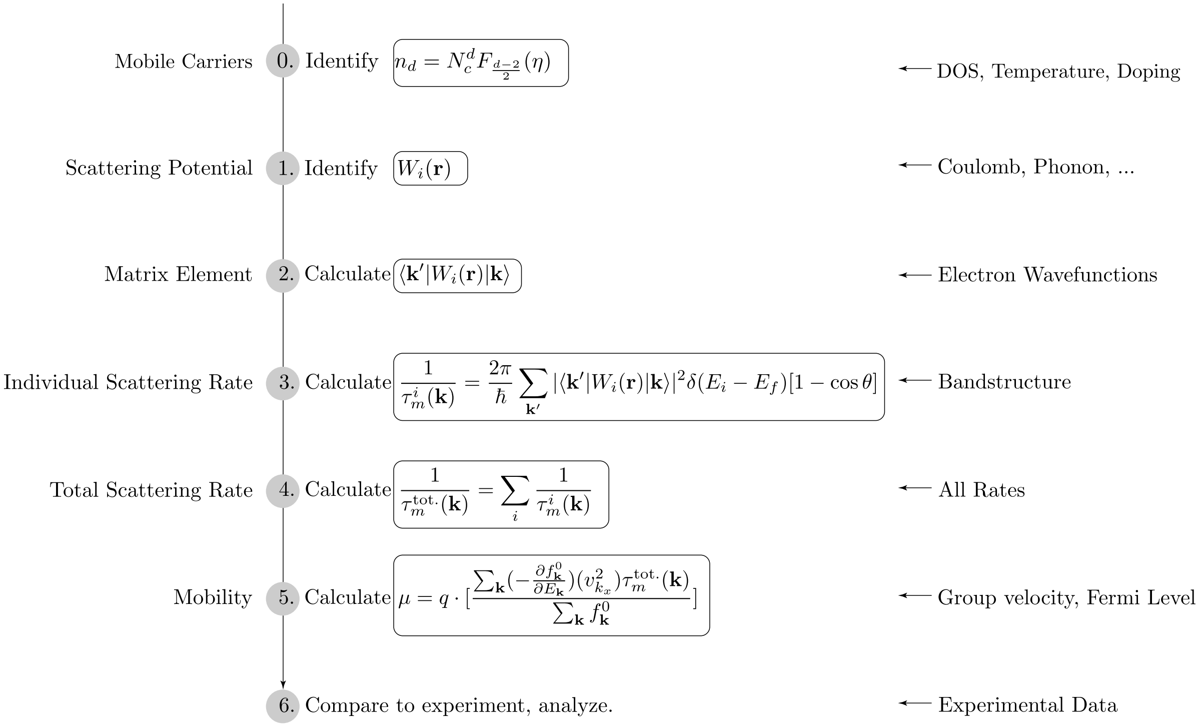
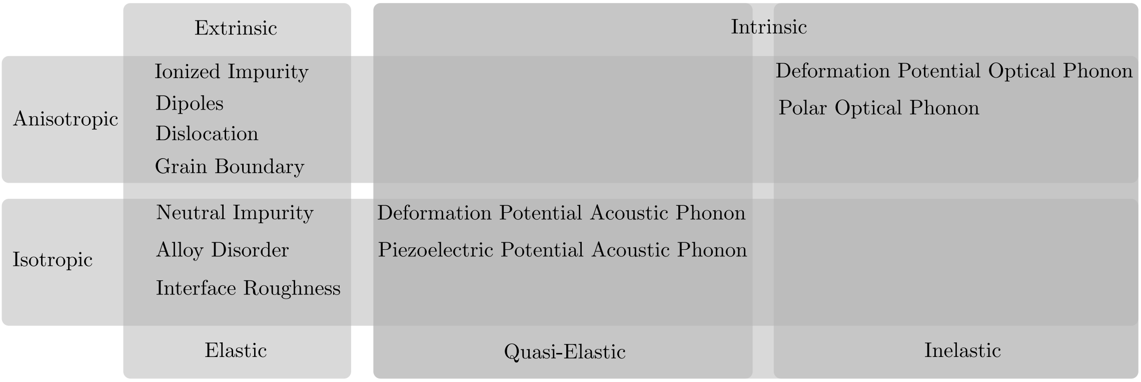

Mobile Carriers
0. Identify $n_d = N_c^d F_{\frac{d-2}{2}}(\eta)$
DOS, Temperature, Doping
Scattering Potential
1. Identify $W_i(\mathbf{r})$
Coulomb, Phonon, ...
Matrix Element
2. Calculate $\langle \mathbf{k}'|W_i(\mathbf{r})|\mathbf{k}\rangle$
Electron Wavefunctions
Individual Scattering Rate
3. Calculate $\frac{1}{\tau_m^i(\mathbf{k})} = \frac{2\pi}{\hbar}\sum_{\mathbf{k}'}|\langle \mathbf{k}'|W_i(\mathbf{r})|\mathbf{k}\rangle|^2\delta(E_i - E_f)[1-\cos\theta]$
Bandstructure
Total Scattering Rate
4. Calculate $\frac{1}{\tau_m^{\mathrm{tot.}}(\mathbf{k})} = \sum_i \frac{1}{\tau_m^i(\mathbf{k})}$
All Rates
Mobility
5. Calculate $\mu = q \cdot [\frac{\sum_{\mathbf{k}}(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}})(v_{k_x}^2)\tau_m^{\mathrm{tot.}}(\mathbf{k})}{\sum_{\mathbf{k}} f_{\mathbf{k}}^0}]$
Group velocity, Fermi Level
6. Compare to experiment, analyze.
Experimental Data

Fig. 23.7 Flow chart for the evaluation of mobility in semiconductors. The last column indicates the information that is necessary to perform the calculations and analysis.

where $g_d(E_{\mathbf{k}})$ is the d-dimensional density of states given in Equation 5.93 and $f_{\mathbf{k}}^0$ is the Fermi–Dirac distribution, which in the non-dimensional form reads $f_{\mathbf{k}}^0(u) = \frac{1}{\exp[u-\eta]+1}$, where $u = \frac{E_{\mathbf{k}}}{k_b T}$. In the RTA solution to the BTE (Equation 21.61), the conductivity always involves the factor $-\frac{\partial f_{\mathbf{k}}^0}{\partial u}$ that integrates to zero for the carrier density, but is responsible for the current by breaking the symmetry in group velocities in the **k**-space. Fig. 23.6 shows the dependence of the density n_d on η. Note that for non-degenerate concentrations the Fermi level is in the energy gap, and $n_d \approx N_c^d e^{\eta} = N_c^d e^{-(E_c - E_F)/k_b T}$, and the carrier density depends exponentially on the temperature. For degenerate concentrations, $n_d \approx N_c^d \cdot \eta^{d/2}/\Gamma((d/2)+1)$, in which the temperature dependence cancels, indicating metallic behavior. These two features have a strong effect on the transport properties of carriers. For 3D and 2D cases with $g_s = 2, g_v = 1$, the densities are[1] $n_{3d} = N_c^{3d} F_{1/2}(\eta)$ and $n_{2d} = N_c^{2d}\ln(1 + e^{\eta})$. For the 2D case, $\eta = (E_F - E_0)/k_b T$ where E_0 is the lowest subband energy.

1. Scattering potential $W_i(\mathbf{r})$ **and screening**: The solution of the Schrödinger equation for the periodic crystal potential $V_0(\mathbf{r})$ results in

[1] If η is needed as a function of the carrier density and temperature, for $d = 3$ this is achieved by inverting the expression $n_{3d} = N_c^{3d} F_{1/2}(\eta)$ by a numerical technique (the Joyce–Dixon approximation): $\eta \simeq \ln(\frac{n_{3d}}{N_c^{3d}}) + \sum_{m=1}^{4} A_m (\frac{n_{3d}}{N_c^{3d}})^m$, where the constants $A_m = 3.536 \times 10^{-1}, -4.950 \times 10^{-3}, 1.484 \times 10^{-4}, -4.426 \times 10^{-6}$ for $m = 1, 2, 3, 4$ respectively.



	Extrinsic	Intrinsic	
	Elastic	Quasi-Elastic	Inelastic
Anisotropic	Ionized Impurity Dipoles Dislocation Grain Boundary		Deformation Potential Optical Phonon Polar Optical Phonon
Isotropic	Neutral Impurity Alloy Disorder Interface Roughness	Deformation Potential Acoustic Phonon Piezoelectric Potential Acoustic Phonon	

Fig. 23.8 Classification of scattering mechanisms with potentials $W_i(\mathbf{r})$ for elastic processes, and $W_{per}(\mathbf{r})e^{\pm i\omega t}$ for inelastic processes.

Bloch eigenfunctions, and electron energy eigenvalues which form the electron bandstructure $E_{\mathbf{k}}$. As discussed in Chapter 13, linear combinations of the Bloch eigenfunctions form the effective mass wavefunctions $\psi(\mathbf{r}) = u(\mathbf{r})C(\mathbf{r})$ where $u(\mathbf{r})$ is the cell-periodic part, and $C(\mathbf{r})$ is the envelope function. Because the potential experienced by electrons is quasi-periodic in doped semiconductors and quantized heterostructures used in practical devices, the effective mass form, and the wavepacket picture associated with it is the correct one for the analysis of transport properties. As discussed in Chapters 14 and 21, the presence of ionized dopants and other defects violates the perfectly periodic potential assumption. But if their density is much lower than the atomic density, it is prudent to assume that the crystal potential is still periodic, with only occasional perturbations of the type $W_i(\mathbf{r})$ separated by many many lattice constants[2]. This key assumption permits the use of Fermi's golden rule developed in Chapter 20 to find the scattering rates between effective mass electronic states due to the perturbing potential.

A heterostructure leads to a rearrangement of the bandstructure itself, modifying the electronic states from the bulk value to effective mass wavefunctions for the particular heterostructure potential. For example, in heterostructure quantum wells there may be a confining potential and bound states in the z-direction, but plane-waves in the $x - y$ plane. In such heterostructures, the perturbations $W_i(\mathbf{r})$ can be due to interface roughness or dopant potentials, but *not the intrinsic heterostructure potential itself, which is periodic and already accounted for in the effective mass solution.*

Potentials that scatter electrons are of two types: **1) static aperiodic potentials** of the type $W_i(\mathbf{r})$ that cause a deviation in the periodic crystal potential that does not change in time, and **2) time-varying periodic potentials** of the type $W_{per}(\mathbf{r})e^{\pm i\omega t}$. As discussed in Chapter 20, the corresponding Fermi's golden rule takes different forms for the two types of perturbation potentials. Static scattering poten-

[2] Therefore, if the density of defects is close to the atomic density, this approximation fails. Then one must go back and re-solve the Schrödinger equation with the scattering potential included directly as part of the crystal potential and not as a perturbation. This may or may not be solvable.

tials of type 1 result in **elastic scattering** of electrons (and holes), in which the energy of the electron before and after scattering is the same $E_{\mathbf{k}} = E_{\mathbf{k}'}$, and only the quasimomentum is changed from $\mathbf{k} \to \mathbf{k}'$, resulting in a scattering rate $S_i(\mathbf{k} \to \mathbf{k}') = \frac{2\pi}{\hbar}|\langle \mathbf{k}'|W_i(\mathbf{r})|\mathbf{k}\rangle|^2 \delta(E_{\mathbf{k}} - E_{\mathbf{k}'})$. Time-varying scatterers of type 2 cause **inelastic scattering**, changing both the quasimomentum $\mathbf{k} \to \mathbf{k}'$ and the energy[3] by $E_{\mathbf{k}'} = E_{\mathbf{k}} \pm \hbar\omega$ where ω is the circular frequency of oscillation, with scattering rates $S_i^{\pm}(\mathbf{k} \to \mathbf{k}') = \frac{2\pi}{\hbar}|\langle \mathbf{k}'|W_{per}(\mathbf{r})|\mathbf{k}\rangle|^2 \delta(E_{\mathbf{k}} - E_{\mathbf{k}'} \pm \hbar\omega)$.

Fig. 23.8 highlights where various scattering mechanisms fall in this classification for semiconductors. Impurities and defects are *in principle* avoidable, and hence extrinsic: they all lead to elastic scattering because of their static potentials. Phonons, whose potentials are periodic and time-varying at finite temperature are *unavoidable, both in principle and in practice*. Phonon scattering is therefore intrinsic, and inelastic. It sets the ultimate upper limit of the attainable electron mobility in any semiconductor. Since very low energy phonons of the acoustic kind do not change the electron energy by a substantial amount, they may be approximated to be quasi-elastic, as discussed later in this chapter.

The scattering potentials of some extrinsic scattering mechanisms such as ionized dopants, and dipoles and dislocations, are long-range. They take the form $W_i(\mathbf{r}) \sim 1/|\mathbf{r}|^n$, which extends over 100s of lattice constants or more. A characteristic feature of scattering of such scattering potentials is they favor small-angle scattering over large-angle scattering, and are thus anisotropic scatterers. For them, the angular factor in the momentum scattering rate $\frac{1}{\tau_m^i(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')[1 - \cos\theta]$ has a large influence. On the other hand, short-range scattering potentials whose potentials $W_i(\mathbf{r})$ are confined to the order of a unit cell size do not favor small or large angles: they are isotropic scatterers. For them, the $\cos\theta$ term sums to zero, making the momentum scattering time equal to the quantum scattering time $\frac{1}{\tau_m^i(\mathbf{k})} = \frac{1}{\tau_q^i(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k} \to \mathbf{k}')$ which was defined in Equation 21.64 and illustrated in Fig. 21.15. A measurement of the macroscopic ratio τ_m/τ_q thus experimentally distinguishes between microscopic long- or short-range scattering potentials[4]. For elastic scattering mechanisms, the energy relaxation time is clearly $\frac{1}{\tau_E} = 0$, and is non-zero for inelastic scattering processes. Mobile electrons (and holes) *screen* and weaken long-range scattering potentials, as discussed next.

Screening: static and dynamic: A time-independent, or static potential due to a point charge q in vacuum is $V(r) = q/(4\pi\epsilon_0 r)$, but inside a semiconductor it is screened to $V(r) = q/(4\pi\epsilon_s r)$. This amounts to a lowering by a factor $\epsilon_s/\epsilon_0 \sim 10 - 20$ for semiconductors, where ϵ_s is the static dielectric constant. This first order screening of Coulomb potentials is caused by the deformation of the valence electron clouds around the atoms of the semiconductor (see Fig. 26.5).

The static dielectric constant of a semiconductor is very crudely related to an effective energy bandgap E_g^{eff}, its valence electron[5] density

[3] A crude analogy is a boat moving in a calm sea going around a small island (a stationary perturbation) changes its direction but not its energy. But in stormy weather, the waves (time-varying perturbation potential) change both the boat's direction and it's energy.

[4] See Chapter 25 for how this is measured in quantum oscillations of the magnetotransport properties.

[5] The valence electron density n_{val} is of the order of 4× the atomic density, for 4 electrons/atom. E_g^{eff} is an averaged gap between the valence and conduction bands over all $\mathbf{k}$ in the Brillouin zone.

n_{val}, and its resulting **plasmon frequency** ω_{pl}, by the relation

$$\frac{\epsilon_s}{\epsilon_0} \approx 1 + (\frac{\hbar\omega_{pl}}{E_g^{eff}})^2, \text{ where } \omega_{pl} = \sqrt{\frac{n_{val}e^2}{\epsilon_0 m_e}}. \tag{23.2}$$

Smaller bandgap semiconductors therefore have higher static dielectric constants and are more "polarizable". As a result, smaller bandgap semiconductors experience weaker Coulomb scattering potentials and as a result have higher mobilities for the same density of impurities.

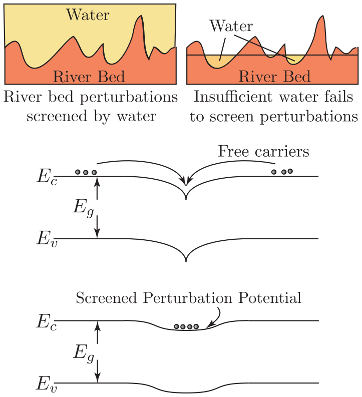

Water
River Bed
River bed perturbations screened by water
Water
River Bed
Insufficient water fails to screen perturbations
Free carriers
E_c
E_g
E_v
Screened Perturbation Potential
E_c
E_g
E_v

Fig. 23.9 Illustration of the phenomena of screening in semiconductors. The top figure gives an analogy of a river bed that is screened by water. The bottom figures show screening by mobile electrons in a semiconductor.

Mobile carriers in the bands of a semiconductor cause additional screening, adding to the static dielectric constant. Fig. 23.9 explains this screening by analogy to a river with a rough bed. A boat moving on the surface is unaffected by the rough bed when the high water level screens the bed. Insufficient water reveals the rough bed and impedes motion of the boat. The analogous screening of a scattering potential in a semiconductor is shown in the energy band diagrams.

This qualitative consideration is given a rigorous quantitative basis by calculating how a dynamic perturbation potential $W_{uns}(\mathbf{r}, t) = W_{uns}(\mathbf{r})e^{-i\omega t}$ written as a Fourier series $\sum_{\mathbf{q}} W_{uns}(\mathbf{q}, \omega) = \sum_{\mathbf{q}} W_{\mathbf{q}} e^{i(\mathbf{q}\cdot\mathbf{r}-\omega t)}$ is modified upon screening by mobile carriers. We simply state the result here[6] and proceed to use it. The screened potential is lowered from the unscreened potential by $\epsilon(\mathbf{q}, \omega)$, the **Lindhard dielectric function**:

$$\boxed{W_{scr}(\mathbf{q}, \omega) = \frac{W_{uns}(\mathbf{q}, \omega)}{\epsilon(\mathbf{q}, \omega)}, \; \epsilon(\mathbf{q}, \omega) = 1 - W_{\mathbf{q}} \sum_{\mathbf{k}} \frac{f_{\mathbf{k}+\mathbf{q}} - f_{\mathbf{k}}}{(E_{\mathbf{k}+\mathbf{q}} - E_{\mathbf{k}}) + \hbar\omega}.} \tag{23.3}$$

Here $f_{\mathbf{k}+\mathbf{q}}$ and $f_{\mathbf{k}}$ are the occupation functions of the Bloch states $\mathbf{k} + \mathbf{q}$ and $\mathbf{k}$, of corresponding energy eigenvalues $E_{\mathbf{k}+\mathbf{q}}$ and $E_{\mathbf{k}}$. $W_{\mathbf{q}}$ is the Fourier component of the spatial part of the perturbation potential, ω is its frequency. Consider a static unscreened scattering potential $W_{uns}(\mathbf{q}, 0)$ with the Fourier component $W_{\mathbf{q}}$. The Lindhard dielectric function for this potential is

$$\begin{aligned} \epsilon(q, 0) &= 1 - W_{\mathbf{q}} \sum_{\mathbf{k}} \frac{f_{\mathbf{k}+\mathbf{q}} - f_{\mathbf{k}}}{E_{\mathbf{k}+\mathbf{q}} - E_{\mathbf{k}}} \approx 1 - W_{\mathbf{q}} \sum_{\mathbf{k}} \frac{\mathbf{q} \cdot \nabla_{\mathbf{k}} f_{\mathbf{k}}^0}{\mathbf{q} \cdot \nabla_{\mathbf{k}} E_{\mathbf{k}}} \\ &\approx 1 + W_{\mathbf{q}} \sum_{\mathbf{k}} (-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}}) \frac{\mathbf{q} \cdot \nabla_{\mathbf{k}} E_{\mathbf{k}}}{\mathbf{q} \cdot \nabla_{\mathbf{k}} E_{\mathbf{k}}} \approx 1 + W_{\mathbf{q}} \sum_{\mathbf{k}} (\frac{1}{k_b T} \frac{\partial f_{\mathbf{k}}^0}{\partial \eta}) \\ &\approx 1 + \frac{W_{\mathbf{q}} L^d}{k_b T} \cdot \frac{\partial (\frac{\sum_{\mathbf{k}} f_{\mathbf{k}}^0}{L^d})}{\partial \eta} \implies \boxed{\epsilon(q, 0) \approx 1 + \frac{W_{\mathbf{q}} L^d}{k_b T} \cdot \frac{\partial n_d}{\partial \eta}.} \end{aligned} \tag{23.4}$$

Because in d-dimensions for parabolic bands $n_d = N_c^d F_{\frac{d}{2}-1}(\eta)$, the factor $\frac{\partial n_d}{\partial \eta} = N_c^d F_{\frac{d}{2}-2}(\eta)$. Several identities encountered in earlier chapters are used here, such as $\nabla_{\mathbf{k}} f_{\mathbf{k}}^0 = -\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}} \nabla_{\mathbf{k}} E_{\mathbf{k}}$, and $-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}} = +\frac{1}{k_b T} \frac{\partial f_{\mathbf{k}}^0}{\partial \eta}$, which is obtained by writing $\eta = \frac{E_F - E_c}{k_b T}$, $u = \frac{E_{\mathbf{k}}}{k_b T}$, and $f_{\mathbf{k}}^0 = \frac{1}{\exp{[u-\eta]}+1}$.

---

[6] The result is obtained by either applying quantum mechanical perturbation theory of Chapters 7 and 20, or from a semi-classical solution of the Boltzmann equation of Chapter 21. A derivation is outlined in Exercise 23.3.

Consider the unscreened potential energy $W_{uns}(r) = \frac{e^2}{4\pi\epsilon_s r}$ of a singly-charged ionized donor[7]. Because the position of the donor is static, $\omega = 0$. The Fourier transform of this unscreened potential in 3D is

$$W(q) = \int \frac{d^3r}{L^3} e^{i\mathbf{q}\cdot\mathbf{r}} \frac{e^2}{4\pi\epsilon_s r} = \frac{e^2}{4\pi\epsilon_s L^3} \int r e^{iqr\cos\theta} \sin\theta dr d\theta d\phi$$

$$= \frac{e^2}{4\pi\epsilon_s L^3} \underbrace{\int_{\phi=0}^{2\pi} d\phi}_{2\pi} \int_{r=0}^{\infty} r dr \underbrace{\int_{\theta=0}^{\pi} d\theta e^{iqr\cos\theta} \sin\theta}_{\frac{2\sin(qr)}{qr}}$$

$$= \frac{e^2}{\epsilon_s q L^3} \int_{r=0}^{\infty} \sin(qr) dr = \frac{e^2}{\epsilon_s q L^3} \lim_{q_0 \to 0} [\int_{r=0}^{\infty} e^{-q_0 r} \sin(qr) dr]$$

$$= \lim_{q_0 \to 0} [\frac{e^2}{\epsilon_s q L^3} \frac{q}{q^2 + q_0^2}] \implies \boxed{W(q) = \frac{e^2}{\epsilon_s q^2 L^3}.} \quad (23.5)$$

Using Equation 23.4, this leads to a static screened dielectric function for the Coulomb potential:

$$\boxed{\epsilon_{3d}(q,0) \approx 1 + \frac{e^2 N_c^{3d} F_{-\frac{1}{2}}(\eta)}{\epsilon_s q^2 k_b T} = 1 + (\frac{q_s^{3d}}{q})^2} \text{ with } q_s^{3d} = \sqrt{\frac{e^2 N_c^{3d} F_{-\frac{1}{2}}(\eta)}{\epsilon_s k_b T}}. \quad (23.6)$$

For non-degnerate conditions, $\eta \ll -1$, and thus $N_c^{3d} F_{-\frac{1}{2}}(\eta) \approx N_c^{3d} e^{\eta} \approx n_{3d}$, the screening wavevector $q_s^{3d} = \sqrt{\frac{e^2 n_{3d}}{\epsilon_s k_b T}}$, and its inverse is known as the **Debye-Huckel screening length** that was first derived in the study of screening in ionic solutions, where $Z \geq 1$ is common. On the other extreme of a degenerate electron concentration $\eta \gg 1$, $F_{-\frac{1}{2}}(\eta) \approx \frac{2\eta^{\frac{1}{2}}}{\sqrt{\pi}}$ leads to a screening wavevector $q_s^{3d} = \sqrt{\frac{3e^2 n_{3d}}{2\epsilon_s (E_F - E_c)}} = 2(\frac{3}{\pi})^{\frac{1}{6}} \sqrt{\frac{n_{3d}^{1/3}}{a_B^\star}} \approx 2\sqrt{\frac{n_{3d}^{1/3}}{a_B^\star}}$, whose inverse is called the **Thomas–Fermi screening length** (see Fig. 23.10). This screening wavevector is independent of temperature, and depends on the (1/6)th power of the 3D carrier concentration and the semiconductor properties via the effective Bohr radius $a_B^\star$.

The screened 3D potential is then obtained as

$$V_{scr}(q) = \frac{V_{uns}(q)}{\epsilon_{3d}(q,0)} = \frac{e^2}{\epsilon_s L^3 [q^2 + (q_s^{3d})^2]}$$

$$\implies V_{scr}(r) = \int \frac{d^3q}{(\frac{2\pi}{L})^3} e^{-i\mathbf{q}\cdot\mathbf{r}} V_{scr}(q) = \frac{e^2}{4\pi\epsilon_s r} e^{-(q_s^{3d} r)}, \quad (23.7)$$

where the effect of screening (Fig. 23.11) is an exponential lowering of the potential, reducing it from a long-range to a short-range force.

Screening for transport in 2D is important for the mobility of 2D electron and hole gases in quantum wells and FETs. Fig. 23.12 shows a point charge a distance z_0 away from the plane in which 2D electrons

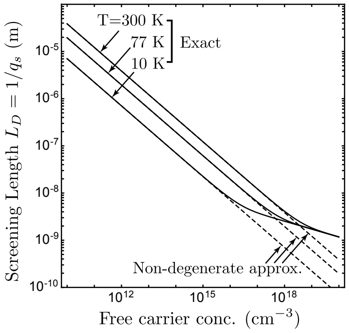


Fig. 23.10 Screening lengths in 3D for $m^\star = 0.2m_e, g_v = 1, \epsilon_s = 10.4\epsilon_0$. The non-degenerate approximation is excellent for $n < 10^{17}/\text{cm}^3$ at 300 and 77 K.

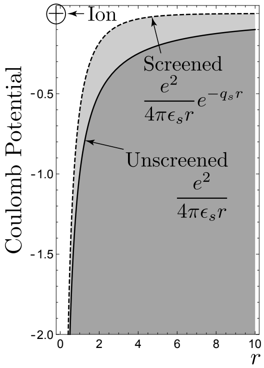


Fig. 23.11 Unscreened and screened attractive potential experienced by an electron around a positively charged ionized donor. This exponentially screened version of the Coulomb potential is called the **Yukawa potential** after Hideki Yukawa who introduced it in the study of pi-mesons in nuclear physics.

---

[7] Note that we are switching for the time being to the notation e for the magnitude of electron charge to avoid confusion with the symbol q which is now used for the electron wavevector in the Fourier expansions and scattering matrix elements. It is typically clear from the context, and if not, a simple dimensional check will suffice.

are confined to move. The unscreened Coulomb potential due to the charge on the plane, and its 2D Fourier transform is

$$W_{uns}(\rho, z_0) = \frac{e^2}{4\pi\epsilon_s\sqrt{\rho^2 + z_0^2}} \implies W_{uns}(q, z_0) = \frac{e^2}{4\pi\epsilon_s L^2}\int \frac{e^{i\mathbf{q}\cdot\rho}\rho d\rho d\phi}{\sqrt{\rho^2 + z_0^2}}$$

$$= \frac{e^2}{4\pi\epsilon_s L^2}\int_{\rho=0}^{\infty} \frac{\rho d\rho}{\sqrt{\rho^2 + z_0^2}} \underbrace{\int_{\phi=0}^{2\pi} d\phi e^{iq\rho\cos\phi}}_{2\pi J_0(q\rho)} = \frac{e^2}{2\epsilon_s L^2} \underbrace{\int_{\rho=0}^{\infty} \frac{\rho J_0(q\rho) d\rho}{\sqrt{\rho^2 + z_0^2}}}_{e^{-q|z_0|}/q}$$

$$\implies \boxed{W_{uns}(q, z_0) = \frac{e^2}{2\epsilon_s q L^2} e^{-q|z_0|},} \text{ and } \boxed{W_{uns}(q, 0) = W_q = \frac{e^2}{2\epsilon_s q L^2},} \tag{23.8}$$

where $J_0(x)$ is the Bessel$-J$ function of order zero. This leads to a 2D screening function $\epsilon_{2d}(q, 0) \approx 1 + \frac{e^2}{2\epsilon_s q k_b T}\frac{\partial n_{2d}}{\partial \eta}$ which is written as

$$\boxed{\epsilon_{2d}(q, 0) \approx 1 + \frac{q_s^{2d}}{q}} \text{ where } \boxed{q_s^{2d} = \frac{e^2 N_c^{2d} F_{-1}(\eta)}{\epsilon_s k_b T} \approx \frac{2}{a_B^\star},} \tag{23.9}$$

and the screened potential is $W_{scr}(q, z_0) = \frac{W_{uns}(q, z_0)}{\epsilon_{2d}(q, 0)}$. For a parabolic band, $N_c^{2d} = \frac{2\pi m_c^\star k_b T}{h^2}$, and $F_{-1}(\eta) = \frac{e^\eta}{1+e^\eta} \approx 1$ makes the 2D screening wavevector $q_s^{2d} = \frac{e^2 m^\star F_{-1}(\eta)}{2\pi\epsilon_s\hbar^2} \approx \frac{2}{a_B^\star}$, where $a_B^\star = \frac{4\pi\epsilon_s\hbar^2}{e^2 m^\star}$ is the effective Bohr radius of the semiconductor. q_s^{2d} is called the 2D **Thomas–Fermi wavevector**. Unlike the exponential screening embodied by the Yukawa potential in 3D, the 2D screened potential in real space goes as a power law. This is indicated in Table 23.1, which summarizes the static and dynamic screening parameters in 2D and 3D free carrier distributions (see Exercise 23.3).

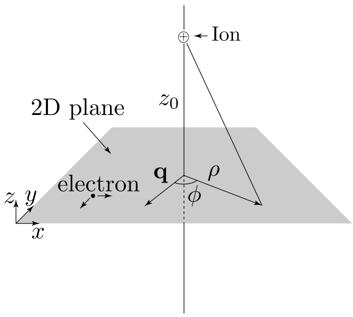


Fig. 23.12 Screening geometry in 2D.

Table 23.1 Screening in semiconductors.

	2D	3D
$\epsilon(q, 0)$	$1 + \frac{q_s^{2d}}{q}$	$1 + (\frac{q_s^{3d}}{q})^2$
$q_s^{non-deg.}$	$q_s^{2d} \approx \frac{2}{a_B^\star}$	$q_s^{3d} \approx \sqrt{\frac{q^2 n_{3d}}{\epsilon_s k_b T}}$
$q_s^{deg.}$	$q_s^{2d} = \frac{2}{a_B^\star}$	$q_s^{3d} = 2\sqrt{\frac{n_{3d}^{\frac{1}{3}}}{a_B^\star}}$
$\epsilon(0, \omega)$	$1 - (\frac{\omega_{pl}^{2d}}{\omega})^2$	$1 - (\frac{\omega_{pl}^{3d}}{\omega})^2$
ω_{pl}	$\sqrt{\frac{e^2 n_{2d} q}{\epsilon_s m^\star}}$	$\sqrt{\frac{e^2 n_{3d}}{\epsilon_s m^\star}}$
$V_{scr.}(r)$	$\frac{e^2}{4\pi\epsilon_s}\frac{q_s^{2d}(1+q_s^{2d}d)}{(q_s^{2d}r)^3}$	$\frac{e^2}{4\pi\epsilon_s r}e^{-q_s^{3d}r}$
$L_s = \frac{1}{q_s}$	$\frac{a_B^\star}{2}$	$\sqrt{\frac{\epsilon_s k_b T}{e^2 n_{3d}}}$

We briefly discuss the limits of screening in space and in time. The Fermi wavelength $\lambda_F = 2\pi/k_F$ of the highest energy mobile electrons sets the lower limit of the length scale over which mobile carriers can screen scattering potentials. The screening dielectric function for purely time-varying potentials (with no spatial variation) are related to the plasmon frequency ω_{pl}, which is obtained by direct evaluation of the Lindhard formula, and listed in Table 23.1. The scattering potentials of phonons vary both in space and in time. Acoustic phonon potentials of low frequency are screened since electrons have enough time to rearrange. Screening of high-frequency polar optical phonons is more difficult, and is limited by the electron scattering by other mechanisms. These will be discussed in the respective sections later in this chapter.

2. Calculating the matrix element $\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle$: As outlined in the flowchart in Fig. 23.7, identifying the scattering potential $W(\mathbf{r})$, and the electron eigenfunctions of the initial and final states allow the direct evaluation of the scattering matrix element:

$$\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle = \int d^3r[\frac{e^{-i\mathbf{k}'\cdot\mathbf{r}}}{\sqrt{V}}u^{\star}_{c\mathbf{k}'}(\mathbf{r})] \cdot W(\mathbf{r}) \cdot [\frac{e^{+i\mathbf{k}\cdot\mathbf{r}}}{\sqrt{V}}u_{c\mathbf{k}}(\mathbf{r})], \qquad (23.10)$$

where $u_{c\mathbf{k}}(\mathbf{r})$ and $u_{c\mathbf{k}'}(\mathbf{r})$ are the periodic parts of the Bloch function. While the periodic parts repeat every unit cell, the scattering potentials $W(\mathbf{r})$ and the envelope functions change only slowly from one unit cell to another, as shown in Fig. 23.13. This difference suggests splitting the integral into a cell-periodic part, and a slowly varying part. This is achieved by writing $\mathbf{r} = \mathbf{R}_i + \mathbf{r}'$ and $V = N_u\Omega_u$ where Ω_u is the volume of a unit cell, and N_u is the number of unit cells. The integral then allows a meaningful split:

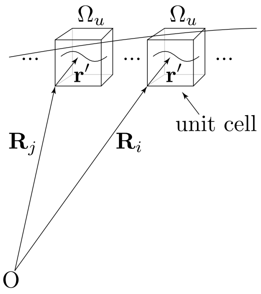


Fig. 23.13 Repeating and non-repeating parts of the matrix element.

$$\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle = \int \underbrace{[\frac{e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}W(\mathbf{r})}{N_u}]}_{\text{slow}} \cdot \underbrace{[u^{\star}_{c\mathbf{k}'}(\mathbf{r})u_{c\mathbf{k}}(\mathbf{r})]}_{\text{periodic}} \cdot \frac{d^3r}{\Omega_u}$$

$$= \underbrace{[\int u^{\star}_{c\mathbf{k}'}(\mathbf{r}')u_{c\mathbf{k}}(\mathbf{r}')\frac{d^3r'}{\Omega_u}]}_{I_{\mathbf{k},\mathbf{k}'}} \cdot [\sum_i \frac{e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{R}_i}W(\mathbf{R}_i)\Omega_u}{N_u\Omega_u}]$$

$$\implies \boxed{\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle = I_{\mathbf{k},\mathbf{k}'} \cdot \int \frac{d^3r}{V}e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}W(\mathbf{r}) = I_{\mathbf{k},\mathbf{k}'}W(\mathbf{q}),} \qquad (23.11)$$

where $I_{\mathbf{k},\mathbf{k}'}$ is the overlap integral of the periodic parts of the Bloch functions, which is typically unity, and $W(\mathbf{q})$ is identified as the Fourier transform of the scattering potential. Note that whenever appropriate this is the Fourier transform of the *screened* potential.

Assume[8] the overlap integral $I_{\mathbf{k},\mathbf{k}'} = 1$. Then, instead of a single scatterer at the origin, if there are N_{imp} *identical* scatterers located at the coordinates $\mathbf{R}_1, \mathbf{R}_2, ...\mathbf{R}_{N_{imp}}$, they have individual scattering potentials $W(\mathbf{r}-\mathbf{R}_1), W(\mathbf{r}-\mathbf{R}_2), ...W(\mathbf{r}-\mathbf{R}_{N_{imp}})$ centered around their locations. In that case, the total scattering potential experienced by an electron is the sum of all the potentials, leading to the squared matrix element[9]:

$$W_{tot}(\mathbf{r}) = W(\mathbf{r}-\mathbf{R}_1) + W(\mathbf{r}-\mathbf{R}_2) + ... + W(\mathbf{r}-\mathbf{R}_{N_{imp}})$$

$$\implies \langle \mathbf{k}'|W_{tot}(\mathbf{r})|\mathbf{k}\rangle = W(\mathbf{q})e^{i\mathbf{q}\cdot\mathbf{R}_1} + W(\mathbf{q})e^{i\mathbf{q}\cdot\mathbf{R}_2} + ...W(\mathbf{q})e^{i\mathbf{q}\cdot\mathbf{R}_{N_{imp}}}$$

$$\implies |\langle \mathbf{k}'|W_{tot}(\mathbf{r})|\mathbf{k}\rangle|^2 = |W(\mathbf{q})|^2 \times$$

$$(e^{i\mathbf{q}\cdot\mathbf{R}_1} + e^{i\mathbf{q}\cdot\mathbf{R}_2} + ... + e^{i\mathbf{q}\cdot\mathbf{R}_{N_{imp}}})(e^{-i\mathbf{q}\cdot\mathbf{R}_1} + e^{-i\mathbf{q}\cdot\mathbf{R}_2} + ... + e^{-i\mathbf{q}\cdot\mathbf{R}_{N_{imp}}})$$

$$= |W(\mathbf{q})|^2 \times (\underbrace{1+1+...1}_{N_{imp}\text{ terms}} + \underbrace{e^{i\mathbf{q}\cdot(\mathbf{R}_1-\mathbf{R}_2)} + e^{i\mathbf{q}\cdot(\mathbf{R}_1-\mathbf{R}_3)} + ...}_{N_{imp}(N_{imp}-1)\text{ terms}})$$

$$\implies \boxed{|\langle \mathbf{k}'|W_{tot}(\mathbf{r})|\mathbf{k}\rangle|^2 = |W(\mathbf{q})|^2 N_{imp}S(\mathbf{q}) \approx N_{imp}|W(\mathbf{q})|^2,} \qquad (23.12)$$

where the structure factor $S(\mathbf{q}) = 1 + \frac{1}{N_{imp}}\sum_{i\neq j} e^{i\mathbf{q}\cdot(\mathbf{R_i}-\mathbf{R_j})}$ is recognized to be the discrete variant of the continuous form encountered in

---

[8] The overlap integral $\int u^{\star}_{c\mathbf{k}'}(\mathbf{r}')u_{c\mathbf{k}}(\mathbf{r}')\frac{d^3r'}{\Omega_u}$ is $= 1$ if the Bloch lattice functions for the $\mathbf{k}$ and $\mathbf{k}'$ states are identical (e.g., the same s-orbital if it is the bottom of the conduction band). This is generally the case if the change in $\mathbf{k}$ is restricted to a small region in the Brillouin zone. When the change in $\mathbf{k}$ is large (for example in intervalley scattering), the factor can deviate from unity and must be evaluated from the bandstructure.

[9] Here, we use the shifting theorem of the Fourier transform, which states that if $W(\mathbf{q})$ is the transform of $W(\mathbf{r})$, then that of $W(\mathbf{r}-\mathbf{R})$ is $W(\mathbf{q})e^{i\mathbf{q}\cdot\mathbf{R}}$.

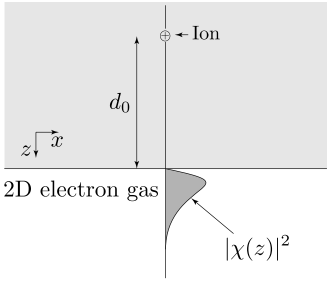


Fig. 23.14 Effect of the finite spread in the 2D electron gas wavefunction on screening and the scattering matrix element.

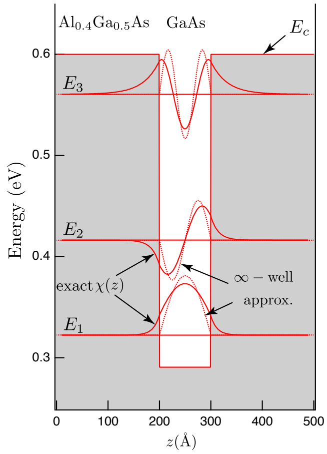


Fig. 23.15 An AlGaAs/GaAs quantum well with square potential profile. The exact and infinite-well approximation of the envelope functions are shown for comparison. Note the increasing penetration of the envelope function into the barrier for increasing bound state energy. The ∞-well approximation for the envelope function becomes less accurate for higher energy bound states. Also see the solved example for such heterostructures in Section 15.6.

Chapter 10, Equation 10.7 in the study of the scattering of X-ray waves from the crystal atoms. Here, the same principle is in action for the scattering of electron waves from defects. If the location of scattering sites are random and uncorrelated, the sum over the random phases $e^{i\mathbf{q}\cdot(\mathbf{R_i}-\mathbf{R_j})}$ vanishes, making $S(\mathbf{q}) = 1$. Then the *square* of the net matrix element increases linearly with the number of scatterers N_{imp}, or in other words, the matrix element itself grows as $\sqrt{N_{imp}}$ which is the nature of **diffusive scattering**. This feature is important in the evaluation of individual scattering rates in the next section. The matrix element is the *amplitude* of scattering from initial state to the final state, whereas its square is the *intensity*. The squared matrix element captures the effect of all the scatterers and their locations through the structure factor. If the identical scatterers were ordered and periodic, then their periodic $W(\mathbf{r})$ must be absorbed in the periodic potential that produced the electron bandstructure in the first place, and there is no scattering of electrons at all: **periodic potentials do not scatter**.

For electron transport in lower-dimensions in quantized heterostructures, evaluation of the matrix element requires the use of appropriate effective mass envelope functions. Because of the importance in electronic and photonic devices, we discuss the general formulation, followed by two widely used analytical approximations for 2D systems. Let $\chi(z)$ be the z-component of the envelope function. Electrons (or holes) are typically not scrunched into a perfect 2D plane with $|\chi(z)|^2 = \delta(z)$ as was depicted in Fig. 23.12. Due to finite barrier heights in heterostructure quantum wells, the envelope wavefunction squared has a finite spread given by $|\chi(z)|^2$ as depicted in Fig. 23.14. Figures 23.15 and 23.16 show the exact and approximate envelope functions for two realistic heterostructures for comparison.

We write the scattering potential $W(\mathbf{r}) = W(\rho, z)$ where $\mathbf{r} = \rho\hat{\rho} + z\hat{z}$ with ρ in the 2D plane and z perpendicular to the plane (see Fig. 23.12). The 2D envelope function is $\frac{e^{i\mathbf{k}\cdot\vec{\rho}}}{\sqrt{L^2}}\chi(z)$, where $\mathbf{k}$ is a 2D wavevector. The matrix element for a 2D electron gas then becomes

$$\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle = \int d^3r[\frac{e^{-i\mathbf{k}'\cdot\vec{\rho}}}{\sqrt{L^2}}\chi_m^\star(z)u_{c\mathbf{k}'}^\star(\mathbf{r})] \cdot W(\mathbf{r}) \cdot [\frac{e^{+i\mathbf{k}\cdot\vec{\rho}}}{\sqrt{L^2}}\chi_n(z)u_{c\mathbf{k}}(\mathbf{r})],$$

$$= I_{\mathbf{k},\mathbf{k}'}\int dz\chi_m^\star(z)\chi_n(z)\int\frac{d^2\rho}{L^2}e^{i\mathbf{q}\cdot\vec{\rho}}W(\rho,z) = \int dz\chi_m^\star(z)\chi_n(z)W(q,z) \tag{23.13}$$

for $I_{\mathbf{k},\mathbf{k}'} = 1$. For a perfect 2D system $\chi_m^\star(z)\chi_n(z) = |\chi(z)|^2 = \delta(z)$, and thus $\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle = W(q,0)$, analogous to the 3D result at $z = 0$. A form-factor $F_{nm}(q)$ is defined for a finite z-spread:

$$\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle = F_{nm}(q)W(q,0), \text{ where } F_{nm}(q) = \frac{\int dz\chi_m^\star(z)\chi_n(z)W(q,z)}{W(q,0)}, \tag{23.14}$$

which by definition takes the value $F_{nm}(q) = 1$ for $|\chi(z)|^2 = \delta(z)$.

Because the squared magnitude of the matrix element is needed in

Table 23.2 Matrix element mathematics for 2D transition rates.

Analytical	δ-fn. well	Fang–Howard–Stern	∞-Well of width L_z
Envelope function for $n=1$: $\chi_1(z)$	-	$\sqrt{\frac{b^3}{2}}ze^{-\frac{bz}{2}}$	$\sqrt{\frac{2}{L_z}}\sin(\frac{\pi}{L_z}z)$
Envelope function for $n=2$: $\chi_2(z)$	-	$\sqrt{\frac{3b^3}{2}}z(1-\frac{b}{3}z)e^{-\frac{bz}{2}}$	$\sqrt{\frac{2}{L_z}}\sin(2\frac{\pi}{L_z}z)$
Envelope function for $n=n$: $\chi_n(z)$	-	-	$\sqrt{\frac{2}{L_z}}\sin(n\frac{\pi}{L_z}z)$
Parameters	-	$b=(\frac{77m^\star q^2 n_{2d}}{16\epsilon_s\hbar^2})^{\frac{1}{3}}$	$\alpha=qL_z,\alpha_z=\beta_zL_z,\alpha_{\parallel}=\beta_{\parallel}L_z$
$\int_0^\infty dz\lvert\chi_1(z)\rvert^4$	-	$\frac{3b}{16}$	$\frac{3}{2L_z}$
$\int_0^\infty dz\lvert\chi_2(z)\rvert^4$	-	$\frac{b}{8}$	$\frac{3}{2L_z}$
$\int_0^\infty dz\lvert\chi_n(z)\rvert^4$	-	-	$\frac{3}{2L_z}$
$P_0=\int\lvert\chi_1(z)\rvert^2e^{-qz}dz$	1	$(\frac{b}{b+q})^3$	$\frac{(1-e^{-\alpha})}{\alpha[1+(\frac{\alpha}{2\pi})^2]}$
$P_1=\int\int\lvert\chi_1(z)\rvert^2\lvert\chi_1(z')\rvert^2e^{-q\lvert z-z'\rvert}dzdz'$	1	$\frac{2(\frac{b}{b+q})^3+3(\frac{b}{b+q})^2+3(\frac{b}{b+q})}{8}$	$\frac{3\alpha^5+20\pi^2\alpha^3+32\pi^4(e^{-\alpha}+\alpha-1)}{\alpha^2(4\pi^2+\alpha^2)^2}$
$F_{12}(q)=\int\chi_1(z)\chi_2(z)e^{-qz}dz$	-	$\frac{\sqrt{3}b^3q}{(b+q)^4}$	$\frac{8\pi^2\alpha(1+e^{-\alpha})}{(\pi^2+\alpha^2)(9\pi^2+\alpha^2)}$
$F_{11}(\beta_z)=\int\lvert\chi_1(z)\rvert^2e^{i\beta_zz}dz$	1	$(\frac{b}{b-i\beta_z})^3$	$\frac{[1-e^{-i\alpha_z})}{i\alpha_z(1-(\frac{\alpha_z}{2\pi})^2]}$
$\lvert F_{11}(\beta_z)\rvert^2$	1	$\frac{b^6}{(b^2+\beta_z^2)^3}$	$\frac{\sin^2(\frac{\alpha_z}{2})}{(\frac{\alpha_z}{2})^2[1-(\frac{\alpha_z}{2\pi})^2]^2}$
$\mathcal{K}=\int_{-\infty}^{+\infty}d\beta_z\lvert F_{11}(\beta_z)\rvert^2$	-	$\frac{3\pi b}{8}$	$\frac{3\pi}{L_z}$
$\mathcal{G}(\beta_{\parallel})=\int_{-\infty}^{+\infty}d\beta_z\frac{\lvert F_{11}(\beta_z)\rvert^2}{\beta_z^2+\beta_{\parallel}^2}$ (Exact)	$\frac{\pi}{\beta_{\parallel}}$	$\frac{\pi}{\beta_{\parallel}}\cdot\frac{(8b^3+9b^2\beta_{\parallel}+3b\beta_{\parallel}^2)}{8(b+\beta_{\parallel})^3}$	$\pi L_z\cdot\frac{3\alpha_{\parallel}^5+20\pi^2\alpha_{\parallel}^3+32\pi^4(e^{-\alpha_{\parallel}}+\alpha_{\parallel}-1)}{\alpha_{\parallel}^3(4\pi^2+\alpha_{\parallel}^2)}$
$\mathcal{G}(\beta_{\parallel})=\int_{-\infty}^{+\infty}d\beta_z\frac{\lvert F_{11}(\beta_z)\rvert^2}{\beta_z^2+\beta_{\parallel}^2}$ (Approx.)	$\frac{\pi}{\beta_{\parallel}}$	$\approx\frac{\pi}{\beta_{\parallel}}\cdot\frac{1}{1+\frac{5}{2}\cdot\frac{\beta_{\parallel}}{b}}$	$\approx\frac{\pi}{\beta_{\parallel}}\cdot\frac{1}{1+\frac{\beta_{\parallel}L_z}{4}}$

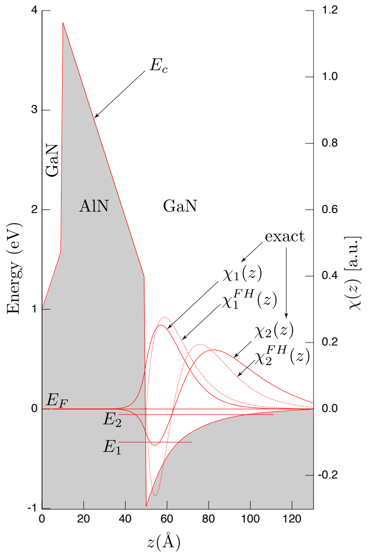


Fig. 23.16 Illustration of the 2DEG at an AlN/GaN heterojunction. The first two exact envelope wavefunctions and the corresponding Fang–Howard variational wavefunctions are shown. Note that the exact envelope functions penetrate the potential barrier, whereas the variational forms do not.

Fermi's golden rule, additional moments of the form factor of the kind

$$|F_{ij}(q)|^2=\int_{z=-\infty}^{\infty}dz\chi_f^\star(z)\chi_i(z)\int dz'\chi_f^\star(z')\chi_i(z')e^{-q|z-z'|} \qquad (23.15)$$

are required. While the exact values may be calculated numerically, analytical evaluation is possible and highly instructive for approximate envelope functions. The generalized form factors and their moments and integrals that frequently are needed in the transport problem are compiled in Table 23.2, and used later in this chapter. Depending on the problem, two analytical envelope functions are commonly used. For square quantum wells (e.g. Fig. 23.15), the ∞-well approximation is used for the envelope functions $\chi_n(z)=\sqrt{\frac{2}{L_z}}\sin(n\frac{\pi}{L_z}z)$, which yields analytical form factors as indicated in the fourth column of Table 23.2. For quantum wells that are more triangular in nature (e.g. Fig. 23.16), the Fang–Howard–Stern variational form $\chi_1(z)=\sqrt{\frac{b^3}{2}}ze^{-bz/2}$ is used for the first subband and $\chi_2(z)=\sqrt{\frac{3b^3}{2}}z(1-\frac{b}{3}z)e^{-\frac{bz}{2}}$ for the second subband. The single variational parameter b for both subbands is dependent on the 2D mobile carrier density, the effective mass and the semiconductor dielectric constant as indicated in the third column of Table 23.2. The barrier height in Fig. 23.16 is larger compared to Fig. 23.15. The barrier height is even larger in Si/SiO$_2$ interfaces, making the Fang–Howard variational wavefunction most accurate for systems of largest barrier height and lowest wavefunction penetration.

Table 23.3 Fang Howard parameters.

$$b = \Big(\frac{33m^\star q^2 n_{2d}}{8\epsilon_s\hbar^2}\Big)^{\frac{1}{3}} = \Big(\frac{33\pi}{2}\frac{n_{2d}}{a_B^\star}\Big)^{\frac{1}{3}} = \Big(\frac{33}{4}\frac{k_F^2}{a_B^\star}\Big)^{\frac{1}{3}}$$
$$P_0 = \frac{b^3}{(b+q)^3}$$
$$P_1 = \frac{8b^3+9b^2q+3bq^2}{8(b+q)^3}$$
$$q_s^{2d} = \frac{e^2}{2\epsilon_s}\frac{dN_s}{dE_F} \approx \frac{2}{a_B^\star}$$
$$\delta_k = \frac{\epsilon_s-\epsilon_b}{\epsilon_s+\epsilon_b}$$

For $q \neq b$,
$$P(z) = \frac{b^3}{(b-q)^3}[e^{-qz} - (a_0 + a_1 z + a_2 z^2)e^{-bz}]$$
$$a_0 = \frac{2q(3b^2+q^2)}{(b+q)^3},\ a_1 = \frac{4bq(b-q)}{(b+q)^2}, a_2 = \frac{q(b-q)^2}{b+q},$$
and for $q = b$,
$$P(z) = \frac{1}{8}[1 + 2bz + 2b^2z^2 + \frac{4}{3}b^3z^3]e^{-bz}.$$

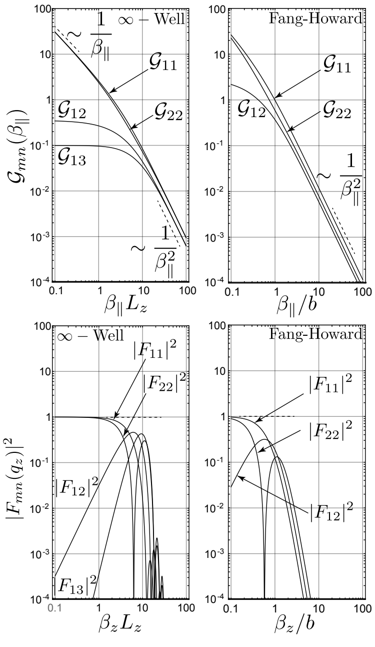


Fig. 23.17 Form factors for square well and the Fang–Howard wavefunctions. Note that the form factors have the property $|F_{nn}(q_z)|^2 \to 1$ for $q_z \to 0$, and that $J_q \to 1/q^2$ for large q.

Because of the finite z-spread of the mobile carriers, screening is slightly modified from the ideal 2D formulation of Equation 23.9. The screened remote Coulomb potential due to a charge at d_0 as shown in Fig. 23.14 for the ground state Fang–Howard–Stern variational function takes the form

$$W_{scr}(q) = \frac{e^2}{2\epsilon_s L^2}\cdot\frac{P_0 e^{-qd_0}}{q + q_s^{2d}(P_1 + \delta_k P_0^2)} \approx \frac{e^2}{2\epsilon_s L^2}\cdot\frac{P_0 e^{-qd_0}}{q + q_s^{2d}P_1} \text{ for } d_0 < 0 \text{ and}$$
$$W_{scr}(q) = \frac{e^2}{2\epsilon_s L^2}\cdot\frac{P(d_0) + \delta_k P_0 e^{-qd_0}}{q + q_s^{2d}(P_1 + \delta_k P_0^2)} \approx \frac{e^2}{2\epsilon_s L^2}\cdot\frac{P(d_0)}{q + q_s^{2d}P_1} \text{ for } d_0 > 0, \tag{23.16}$$

where the parameters are defined in Table 23.3, and P_0, P_1 also appear in Table 23.2. Physically the ion is embedded *inside* the mobile electron (or hole) gas for $d_0 > 0$ and is "remote" for $d_0 < 0$ as indicated in Fig. 23.14. $\delta_k = \frac{\epsilon_s - \epsilon_b}{\epsilon_s + \epsilon_b}$ is a dielectric mismatch factor with ϵ_b the dielectric constant of the barrier layer, and ϵ_s that of the semiconductor. The approximate screened potentials take the form $W_{scr}(q) \approx W_{uns}(q) \times \frac{P_0}{(1+\frac{q_s^{2d}}{q}P_1)}$ for $\delta_k \approx 0$, i.e. when the dielectric mismatch is negligible.

An electron confined in 2D has an in-plane momentum $\mathbf{k} = k_\parallel \hat{\rho}$. But wave-like scatterers such as phonons have an out-of plane z-component of momentum in the form $W(r) \sim e^{i(\beta_\parallel \rho + \beta_z z)}$, leading to the factors

$$F_{nm}(\beta_z) = \int \chi_n(z)\chi_m(z)e^{i\beta_z z}dz, \text{ and } \mathcal{G}_{mn}(\beta_\parallel) = \int_{-\infty}^{+\infty} d\beta_z \frac{|F_{mn}(\beta_z)|^2}{\beta_z^2 + \beta_\parallel^2} \tag{23.17}$$

for the ∞-well and the Fang–Howard envelope functions. Fig. 23.17 shows how the squared form factor and one of its moments depend on the wavevector $\beta_\parallel$ and β_z for the two analytical envelope functions. These integral factors (also listed in Table 23.2) will appear in the evaluation of various scattering rates ranging from impurity scattering to phonon scattering. They will be used later in the chapter to evaluate the scattering rates and mobilities for 2D carrier systems.

While the quantitative details of the screened squared matrix elements, which take the general form $|\langle \mathbf{k}'|W_{tot}(\mathbf{r})|\mathbf{k}\rangle|^2 \sim N_{imp}|W(\mathbf{q})|^2$ lead to various expressions for different geometries and dimensions (and can appear involved), their interpretation and usage is straightforward. The squared matrix element is a function of $\mathbf{q}$, where $\hbar\mathbf{q} = \hbar\mathbf{k} - \hbar\mathbf{k}'$ is the change in the electron momentum due to the scattering event, and the more scatterers there are, the more intense is the scattering. The scattering matrix element is necessary in the next step: to evaluate the scattering rate (or frequency) due to each type of scatterer.

3. **Evaluating the individual scattering rates:** In Chapter 21, Equation 21.68, using the RTA solution of the BTE we found that the momentum relaxation rate of mobile electrons (or holes) in state $\mathbf{k}$ is

$\frac{1}{\tau_m(\mathbf{k})} = \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}')(1 - \cos\theta)$, where θ is the angle between in the initial and final electron wavevectors $\mathbf{k}$ and $\mathbf{k}'$. The individual scattering rate $S(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar}|\langle\mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle|^2\delta(E_i - E_f)$ was obtained from Fermi's golden rule, where E_i is the initial electron energy before the scattering, and E_f the final electron energy after the scattering event. Using the squared matrix element evaluated in Equation 23.12 in the last subsection, the momentum scattering for mobile carriers becomes

$$\frac{1}{\tau_m(\mathbf{k})} = \frac{2\pi}{\hbar}\sum_{\mathbf{k}'}|\langle\mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle|^2\delta(E_i - E_f)[1 - \cos\theta]$$
$$= \frac{2\pi}{\hbar}N_{imp}\sum_{\mathbf{k}'}|I_{\mathbf{k},\mathbf{k}'}|^2 \cdot |W(q)|^2 \cdot \delta(E_i - E_f) \cdot [1 - \cos\theta]. \qquad (23.18)$$

For a general scattering event, conservation of momentum enforces $\hbar\mathbf{k}' = \hbar\mathbf{k} \pm \hbar\mathbf{q}$, and conservation of energy enforces $E_{\mathbf{k}'} = E_{\mathbf{k}} \pm \hbar\omega_{\mathbf{q}}$. Here, $\hbar\mathbf{q}$ is the momentum of the object that scatters the electron (such as a phonon), $\hbar\omega_{\mathbf{q}}$ is the energy of the scatterer, and $E_i = E_{\mathbf{k}}$ and $E_f = E_{\mathbf{k}'}$. To proceed further in computing the rates, the bandstructure of the semiconductor, as well as the energy-momentum dispersions of the scatterers (e.g. phonons) are needed. In the following discussion we consider a parabolic bandstructure for electrons, with the understanding that related non-parabolic (including linear) bandstructures can be evaluated using the same principles. In Fig. 23.8 the scattering mechanisms were classified into elastic and inelastic. We proceed using this classification by discussing elastic and inelastic scattering for 3D, 2D, and 1D carriers.

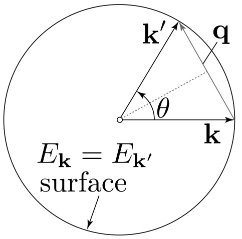


Fig. 23.18 Elastic scattering geometry.

Elastic scattering: Scattering from defects is elastic, ensuring $\hbar\mathbf{q} \approx 0$ and $\hbar\omega_{\mathbf{q}} \approx 0$, making $|\mathbf{k}| = |\mathbf{k}'|$ and $E_{\mathbf{k}} = E_{\mathbf{k}'}$. For parabolic bandstructure with $E_{\mathbf{k}} = \frac{\hbar^2}{2m^\star}|\mathbf{k}|^2$, from the geometry in Fig. 23.18 we obtain $|\mathbf{q}| = 2|\mathbf{k}|\sin\frac{\theta}{2} = 2|\mathbf{k}'|\sin\frac{\theta}{2}$, and therefore $1 - \cos\theta = \frac{q^2}{2k^2}$. For $|I_{\mathbf{k},\mathbf{k}'}| = 1$, the momentum relaxation rate then becomes

$$\frac{1}{\tau_m(\mathbf{k})} = \frac{4\pi m^\star N_{imp}}{\hbar^3}\sum_{\mathbf{k}'}\frac{|\mathbf{k}' - \mathbf{k}|^2}{2k^2} \cdot |W(\mathbf{k}' - \mathbf{k})|^2 \cdot \delta[(k')^2 - k^2]$$
$$= \frac{4N_{imp}\pi m^\star}{\hbar^3}\int\frac{(k')^2\sin\theta dk' d\theta d\phi}{(\frac{2\pi}{L})^3}(1 - \cos\theta) \cdot |W(\mathbf{k}' - \mathbf{k})|^2 \cdot \delta[(k')^2 - k^2]$$
$$= \frac{N_{imp}L^3m^\star}{\pi\hbar^3}\int_{\theta=0}^{\pi} d\theta\sin\theta(1 - \cos\theta)\underbrace{\int_{q=0}^{2k}\frac{dq \cdot |W(q)|^2 \cdot q^2}{[2\sin(\frac{\theta}{2})]^3} \cdot \delta[(\frac{q}{2k\sin(\frac{\theta}{2})})^2 - k^2]}_{=\frac{k}{2}|W(2k\sin\frac{\theta}{2})|^2 \text{ by using } \delta[(k')^2 - k^2] = \frac{\delta[k' - k]}{2k}}$$
$$\boxed{\frac{1}{\tau_m(\mathbf{k})} = \frac{N_{imp}L^3m^\star k}{2\pi\hbar^3}\int_0^{\pi} d\theta \cdot \sin\theta \cdot (1 - \cos\theta) \cdot |W(2k\sin\frac{\theta}{2})|^2,} \qquad (23.19)$$

which is converted from the integral over angles to an integral over wavevectors by using $q = 2k \sin \frac{\theta}{2}$ to get

$$\boxed{\frac{1}{\tau_m(\mathbf{k})} = \frac{N_{imp} L^3 m^\star}{4\pi (\hbar k)^3} \int_0^{2k} dq \cdot q^3 \cdot |W(q)|^2,} \tag{23.20}$$

which is a form we will use frequently to compute scattering rates. The corresponding results for momentum relaxation rate for elastic scattering for 2D carriers is obtained in an identical manner by summing over the 2D **k**-space:

$$\boxed{\frac{1}{\tau_m(\mathbf{k})} = \frac{N_{imp} L^2 m^\star}{2\pi \hbar^3} \int_0^{2\pi} d\theta \cdot (1 - \cos\theta) \cdot |W(2k \sin \frac{\theta}{2})|^2.}$$

$$\boxed{\frac{1}{\tau_m(\mathbf{k})} = \frac{N_{imp} L^2 m^\star}{2\pi (\hbar k)^3} \int_0^{2k} dq \frac{q^2 |W(q)|^2}{\sqrt{1 - (\frac{q}{2k})^2}}.} \tag{23.21}$$

In the evaluation of the above integral for 2D envelope wavefunctions with form factors $P_0(q_\|)$ and $P_1(q_\|)$, variants of the following generalized dimensionless integral appear frequently:

$$I(r, s, t) = \int_0^1 du \frac{u^r e^{-su} P_0^2(u)}{\sqrt{1 - u^2} [u + t P_1(u)]^2}, \tag{23.22}$$

where the form factors implicitly depend on the 2D carrier density n_{2d} via the dependence of $q_\|$ on the Fermi wavevector $k_F = \sqrt{2\pi n_{2d}}$ and the variational parameters b. In general, this integral has to be evaluated numerically. In the special case of 2D carriers are tightly confined with no spread in the z-direction with $\chi^2(z) = \delta(z)$, the form factors become unity: $P_0(u) = P_1(u) = 1$, and the integral can be evaluated analytically exactly for some cases, and approximately for other cases as listed in Table 23.4. These will be useful, for example, in the evaluation of screened remote impurity scattering in modulation doped heterostructures and in interface roughness scattering.

Table 23.4 2DEG Integrals.

$I(r,s,t) = \int_0^1 du \frac{u^r e^{-su} P_0^2(u)}{\sqrt{1-u^2}[u+tP_1(u)]^2}$ For $P_0(u) = P_1(u) = 1$
$I(0,0,t) = \frac{\sqrt{1-t^2} + t^2 \ln(\frac{1-\sqrt{1-t^2}}{t})}{t(1-t^2)^{\frac{3}{2}}}$ $t \to 0 \approx \frac{1}{t}$ and $t \to \infty \approx \frac{1}{t^2}$ $I(0,0,t) \approx \frac{1}{t(1+t)}$
$I(1,0,t) = \frac{1}{t^2-1} - \frac{\mathrm{ArcSec}(t)}{(t^2-1)^{\frac{3}{2}}}$ $t \to 0 \approx -\ln(t)$ and $t \to \infty \approx \frac{1}{t^2}$ $I(1,0,t) \approx \frac{-\ln(t)}{1-t^2}$
$I(2,0,t) = \frac{\pi}{2} - \frac{t}{t^2-1} - \frac{t(t^2-2)\mathrm{ArcSec}(t)}{(t^2-1)^{\frac{3}{2}}}$ $t \to 0 \approx \frac{\pi}{2}$ and $t \to \infty \approx \frac{1}{t^2}$ $I(2,0,t) \approx \frac{\frac{\pi}{2}}{1+t^2}$
$I(3,0,t) = 1 - \pi t + \frac{t^2}{t^2-1} + \frac{t^2(2t^2-3)\mathrm{ArcSec}(t)}{(t^2-1)^{\frac{3}{2}}}$ $t \to 0 \approx 1$ and $t \to \infty \approx \frac{1}{t^2}$ $I(3,0,t) \approx \frac{1}{1+t^2}$

Inelastic scattering: The primary inelastic scattering mechanisms are due to phonons. In Chapter 22, we identified how energy and momentum conservation relations restrict the ranges of phonon wavevectors β and the scattering angles for the electrons. Since the energy change of electrons upon scattering by acoustic phonons is negligible, the scattering is almost elastic. Therefore, the formulation described in Equations 23.19 and 23.21 apply to an excellent approximation for most practical cases, even at low temperatures. A momentum relaxation time is not strictly defined for strongly inelastic scattering events due to optical phonons because the BTE cannot be linearized under general conditions, and solutions must be sought numerically. However, increasingly accurate estimates of momentum relaxation rates due to optical phonon scattering can be obtained under special conditions, which surprisingly hold for a large number of semiconductors. We will take this approach in Section 23.9 on phonon scattering.

4. Evaluating the total scattering rate: If the scattering events occur independent of each other, they occur in parallel and the rates must add. In this approximation, the fastest scattering rates are the dominant ones, which is captured by the Matheissen's rule: $\frac{1}{\tau_m^{\text{tot.}}(\mathbf{k})} = \sum_i \frac{1}{\tau_m^i(\mathbf{k})}$ for each scattering mechanism i. This approximation is an excellent one for all elastic scattering events, but is not strictly accurate for inelastic scattering events, because of the difficulty of defining a momentum relaxation time via the BTE. Nevertheless, the ease of its use, and also because even with the crude approximation the theoretically calculated transport properties such as electron mobilities accurately capture the experimentally measured values over a large range of temperatures and conditions is the reason why Matheissen's rule continues to be used to explain and analyze experimental data.

Because inelastic scattering by high-energy optical phonons connects $f(E) \leftrightarrow f(E \pm \hbar\omega_{LO})$, for increased accuracy one must resort back to the BTE, and solve the coupled forms of the equations for inelastic scattering. This is done by using the optical phonon absorption rate $S_{abs}(k)$ and emission rates $S_{em}(k)$, and ensuring electron density conservation in scattering events. Techniques of varying computational resources exist for efficient numerical evaluation, ranging from Rode's iterative methods, to stochastic Monte Carlo methods. In this chapter, we primarily discuss Matheissen's rule and in the Exercises, outline how the numerical techniques are employed.

5. Calculating the mobility: A generalized formula for mobility was derived in Chapter 21, Equation 21.63. We now re-derive and discuss the general expression for the drift mobility of carriers of arbitrary degeneracy confined in d spatial dimensions, where d may be 1, 2 or 3; for $d = 0$, the carrier in principle does not move in response to a field[10]. Let the electric field be applied along the i^{th} spatial dimension, ($\mathbf{E} = E_i \mathbf{i}$) and the magnetic field $\mathbf{B} = 0$. We assume an isotropic effective mass $m^\star$. Starting from the Boltzmann equation for the distribution function of carriers $f(k, r, t)$, and using the relaxation-time approximation solution, we write the distribution function as

$$f(\mathbf{k}) = f_0(\mathbf{k}) + qF_i \tau(k) v_i \frac{\partial f_0}{\partial \varepsilon}, \tag{23.23}$$

where $\tau(k)$ is the momentum relaxation time and v_i is the velocity of carriers in the i^{th} direction in response to the field. The total number of carriers per unit "volume" in the d-dimensional space is

$$n = \int \frac{d^d k}{(2\pi)^d} f(k) = \int d\varepsilon f(\varepsilon) g_d(\varepsilon), \tag{23.24}$$

where the generalized d-dimensional DOS expressed in terms of the

[10] The 0-D case of quantum dots requires the carriers to tunnel in or out, the topic of Chapter 24. No band-transport is possible in this case.

energy ε of carriers is given by

$$g_d(\varepsilon) = \frac{1}{2^{d-1}\pi^{\frac{d}{2}}\Gamma(\frac{d}{2})}(\frac{2m^\star}{\hbar^2})^{\frac{d}{2}}\varepsilon^{\frac{d}{2}-1}. \tag{23.25}$$

Here $\hbar$ is the reduced Planck's constant and $\Gamma(...)$ is the gamma function. Using this, and the parabolic dispersion we can switch between the k-space and energy-space. The current in response to the electric field along the i^{th} direction is

$$\mathbf{J} = 2q\int \frac{d^d k}{(2\pi)^d}\mathbf{v}f(\mathbf{k}). \tag{23.26}$$

Using the distribution function from the solution of the BTE, we see that the f_0 term integrates out to zero, and only the second term contributes to a current. For a particle moving in d-dimensions the total kinetic energy ε is related to the average squared velocity $\langle v_i^2 \rangle$ along *one* direction by the expression $\langle v_i^2 \rangle = 2\varepsilon/dm^\star$. Using this result, we re-write the current as

$$J_i = qn\underbrace{(-\frac{2e}{dm^\star}\frac{\int d\varepsilon\tau_m\varepsilon^{\frac{d}{2}}\frac{\partial f_0}{\partial\varepsilon}}{\int d\varepsilon f_0(\varepsilon)\varepsilon^{\frac{d}{2}-1}})}_{\mu_d}F_i, \tag{23.27}$$

where the mobility in the d-dimensional case is denoted by the underbrace. τ_m, the momentum relaxation time due to scattering events calculated in the Born approximation by Fermi's golden rule using the scattering potential, turns out to depend on the energy of the mobile carrier and the temperature. Let us assume that it is possible to split off the energy dependence of the relaxation time in the form

$$\tau_m = \tau_0\left(\frac{\varepsilon}{k_b T}\right)^r, \tag{23.28}$$

where τ_0 does not depend upon the energy of the carriers. Using this, and the fact that $f_0(\varepsilon) \to 0$ as $\varepsilon \to \infty$ and $\varepsilon^m \to 0$ as $\varepsilon \to 0$, the expression for mobility can be converted by an integration by parts to

$$\mu_d = \frac{q\tau_0}{m^\star}\cdot\left(\frac{\Gamma(\frac{d}{2}+r+1)}{\Gamma(\frac{d}{2}+1)}\right)\cdot\left(\frac{F_{\frac{d}{2}+r-1}(\eta)}{F_{\frac{d}{2}-1}(\eta)}\right), \tag{23.29}$$

where $F_j(\eta)$ are the traditional Fermi–Dirac integrals of the j^{th} order defined as a function of $\eta = (E_F - E_c)/(k_b T)$ as

$$F_j(\eta) = \frac{1}{\Gamma(j+1)}\int_0^\infty dx\frac{x^j}{1+e^{x-\eta}}. \tag{23.30}$$

Equation 23.29 may be viewed as a *generalized* formula for mobility of carriers in d-dimensional space. Note that this is a *general* expression that holds true for an *arbitrary* degeneracy of carriers that are confined in *arbitrary* (d) dimensions. We now proceed to use this form of the expression for determining the mobility for two extreme cases. The strongly non degenerate ("ND") case, where $\eta \ll -1$ and the strongly degenerate ("D") case, where $\eta \gg +1$.

For the non-degenerate case, the Fermi integrals can be shown to reduce to $F_j(\eta) \approx e^{\eta}$. This reduces the expression for mobility to the simple form

$$\mu_d^{ND} \approx \frac{q\tau_0}{m^\star}\left(\frac{\Gamma(\frac{d}{2}+r+1)}{\Gamma(\frac{d}{2}+1)}\right). \tag{23.31}$$

For the strongly degenerate case, we make another approximation of the Fermi–Dirac integral. We re-write it as

$$F_j(\eta) = \frac{e^{\eta}}{\Gamma(j+1)}[\int_0^{\eta} dx \frac{x^j}{e^{\eta}+e^x} + \int_{\eta}^{\infty} dx \frac{x^j}{e^{\eta}+e^x}]. \tag{23.32}$$

In the first integral, $e^x \ll e^{\eta}$ and in the second integral, $e^x \gg e^{\eta}$ since $\eta \gg 1$. Using this and retaining only the *leading* power of η, the Fermi–Dirac integral can be approximated as $F_j(\eta) \approx \eta^{j+1}/\Gamma(j+2)$. Further, $\eta = \varepsilon_F/k_bT$ where ε_F is the Fermi energy that is known for the degenerate case if one knows *only* the carrier density. So the expression for degenerate carrier mobility finally reduces to the simple form

$$\mu_d^{D} \approx \frac{q\tau_0}{m^\star}\left(\frac{E_F-E_c}{k_bT}\right)^r. \tag{23.33}$$

The validity of the degenerate and non-degenerate limits rests on the accuracy of the approximations made to the Fermi–Dirac integrals. For strong degeneracy and non-degeneracy, the approximations for the three-dimensional case are shown with the exact Fermi–Dirac integrals in Fig. 23.6. We summarize the results now: the generalized carrier mobility is

$$\mu = q\cdot[\frac{\sum_{\mathbf{k}}(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}})(v_{k_x}^2)\tau_m(\mathbf{k})}{\sum_{\mathbf{k}} f_{\mathbf{k}}^0}] = q\cdot[\frac{\int dE_{\mathbf{k}} g(E_{\mathbf{k}})(\frac{f_{\mathbf{k}}^0(1-f_{\mathbf{k}}^0)}{k_bT})(\frac{2E_{\mathbf{k}}}{m^\star d})\tau_m(E_{\mathbf{k}})}{\int dE_{\mathbf{k}} g(E_{\mathbf{k}}) f_{\mathbf{k}}^0}]$$

$$\implies \boxed{\mu = \frac{2q}{m^\star d k_b T}\cdot[\frac{\int dE_{\mathbf{k}} E_{\mathbf{k}}^{\frac{d}{2}} f_{\mathbf{k}}^0(1-f_{\mathbf{k}}^0)\tau_m(E_{\mathbf{k}})}{\int dE_{\mathbf{k}} E_{\mathbf{k}}^{\frac{d}{2}-1} f_{\mathbf{k}}^0}],} \tag{23.34}$$

and if the scattering rate can be expressed as $\tau_m(\mathbf{k}) = \tau_0(\frac{E_{\mathbf{k}}}{k_bT})^r$,

$$\mu = \frac{2q}{m^\star d}\cdot[\frac{\int dE_{\mathbf{k}}(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}})E_{\mathbf{k}}^{\frac{d}{2}}\tau_0(\frac{E_{\mathbf{k}}}{k_bT})^r}{\int dE_{\mathbf{k}} E_{\mathbf{k}}^{\frac{d}{2}-1} f_{\mathbf{k}}^0}] = \frac{2q\tau_0}{m^\star d(k_bT)^r}\cdot[\frac{\int dE_{\mathbf{k}}(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}})E_{\mathbf{k}}^{\frac{d}{2}+r}}{\int dE_{\mathbf{k}} E_{\mathbf{k}}^{\frac{d}{2}-1} f_{\mathbf{k}}^0}]$$

$$= \frac{2q\tau_0(\frac{d}{2}+r)}{m^\star d(k_bT)^r}\cdot[\frac{\int dE_{\mathbf{k}} E_{\mathbf{k}}^{\frac{d}{2}+r-1} f_{\mathbf{k}}^0}{\int dE_{\mathbf{k}} E_{\mathbf{k}}^{\frac{d}{2}-1} f_{\mathbf{k}}^0}] \implies \boxed{\mu = \frac{q\tau_0}{m^\star}\cdot\frac{\Gamma(\frac{d}{2}+r+1)}{\Gamma(\frac{d}{2}+1)}\cdot\frac{F_{\frac{d}{2}+r-1}(\eta)}{F_{\frac{d}{2}-1}(\eta)},} \tag{23.35}$$

where we have used integration by parts to get $\int_{E_{\mathbf{k}}=0}^{\infty} dE_{\mathbf{k}} (-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}}) E_{\mathbf{k}}^{\frac{d}{2}+r} = [-E_{\mathbf{k}}^{\frac{d}{2}+r} f_{\mathbf{k}}^0]|_{E_{\mathbf{k}=0}}^{\infty} + (\frac{d}{2}+r) \int_{E_{\mathbf{k}}=0}^{\infty} dE_{\mathbf{k}} E_{\mathbf{k}}^{\frac{d}{2}+r-1} f_{\mathbf{k}}^0$, in which the first term on the RHS is zero at both $E_{\mathbf{k}} = 0$, and at high energies as $E_{\mathbf{k}} \to \infty$, the Fermi–Dirac distribution $f_{\mathbf{k}}^0(E_{\mathbf{k}}) \to 0$ faster than any polynomial power of $E_{\mathbf{k}}$.

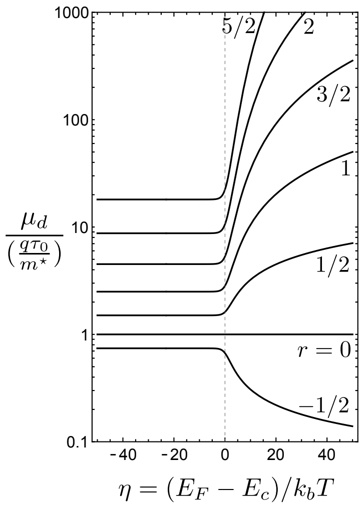


Fig. 23.19 The degeneracy and dimensionality factor for mobility for 3D carriers ($d = 3$) when the scattering rate can be expressed as $\tau_{\mathbf{k}} = \tau_0 (\frac{E_{\mathbf{k}}}{k_bT})^r$.

For non-degenerate electron distributions with $\eta \ll -1$, $F_j(\eta) \approx e^{\eta}$, and for degenerate distributions with $\eta \gg +1$, $F_j(\eta) \approx \eta^{j+1}/\Gamma(j+2)$, which simplifies the expression for mobility to

$$\boxed{\mu_d^{ND} \approx \frac{q\tau_0}{m^\star} \cdot \frac{\Gamma(\frac{d}{2}+r+1)}{\Gamma(\frac{d}{2}+1)}} \text{ and } \boxed{\mu_d^{Deg.} \approx \frac{q\tau_0}{m^\star} \cdot \eta^r = \frac{q}{m^\star}\tau(E_F).} \tag{23.36}$$

The above results are very handy since they allow reasonably accurate estimates of carrier mobilities to be obtained analytically. In particular, the degenerate approximation $\mu_d^{Deg.} \approx \frac{q}{m^\star}\tau(E_F)$ is a remarkable simplification: because of Pauli blocking for degenerate carriers, all transport occurs near the Fermi surface, and thus the mobility can be computed directly by using the net scattering rate evaluated at the Fermi energy[11]. Fig. 23.19 shows the dependence of the dimensionless factor $\mu/((q\tau_0)/m^\star)$ of Equation 23.35 on the degeneracy of the carrier distribution η for 3D carriers. It highlights both the non-degenerate and degenerate limits found in Equation 23.36. The mobility becomes independent of η and depends only on the power r characteristic of the scattering mechanism for $\eta << -1$. For the degenerate extreme, $\eta >> +1$ and a power-law dependence η^r again characterized by r corresponding to the scattering mechanism is expected to appear in the experimentally measured mobility.

6. Comparing with experiment: After evaluating steps $1-5$ of this section, the computed mobility and its dependence on parameters such as the temperature or the carrier concentration are ready to be compared with experimental data such as those discussed in Section 23.1. This enables the identification of the dominant scattering mechanisms at the desired physical condition of temperature or carrier concentration[12]. In the next sections we evaluate the scattering rates explicitly, and then illustrate the comparison with experiment.

23.3 Point defect scattering

3D carriers: Scattering potentials of point defects are highly localized around the defect site in the crystal. The simplest model for the scattering potential is the 3D Dirac delta function. For a point defect located at $\mathbf{r} = \mathbf{r}_i$, the scattering potential is written as $W_i(\mathbf{r}) = W_0\Omega_0\delta(\mathbf{r}-\mathbf{r}_i)$ where Ω_0 is the volume occupied by the defect, and W_0 is the strength

[11] Actually the degenerate approximation also holds if $\tau(E_{\mathbf{k}})$ cannot be split into a power law dependence on energy. This is why it is a rather powerful result. It is specially useful for 2D electron and hole gases.

[12] In the development of semiconductor materials for device applications, this is an important exercise since it provides guidance to improve mobilities by identifying the defects that are the most important. This exercise of comparison also helps highlight how the *relative* importance of scattering mechanisms change, for example with carrier concentrations. This is important in the dynamical properties where such densities *in the same device* can vary over orders of magnitude while in operation.

in units of energy. Note that for 3D, $\delta(\mathbf{r}-\mathbf{r}_i)$ has units of 1/Volume, and thus the scattering potential has units of energy, as it always should. Using the recipe outlined in Section 23.2, we first evaluate the squared matrix element, and then use Equation 23.20 to obtain

$$W_i(\mathbf{r}) = W_0\Omega_0\delta(\mathbf{r}-\mathbf{r}_i) \implies W(\mathbf{q}) = \frac{W_0\Omega_0}{L^3}e^{i\mathbf{q}\cdot\mathbf{r}_i}$$

$$|W_{tot.}(q)|^2 = N_{imp}\frac{W_0^2\Omega_0^2}{L^6} \implies \frac{1}{\tau_m^{pt}(\mathbf{k})} = \frac{N_{imp}L^3 m^\star}{4\pi(\hbar k)^3}\int_{q=0}^{2k} dq\cdot q^3\cdot|W(q)|^2$$

$$= \frac{N_{imp}}{L^3}\frac{m^\star W_0^2\Omega_0^2}{4\pi(\hbar k)^3}\int_{q=0}^{2k} dq\cdot q^3 \implies \boxed{\frac{1}{\tau_m^{pt}(\mathbf{k})} = n_{imp}\frac{m^\star W_0^2\Omega_0^2}{\pi\hbar^3}k.} \tag{23.37}$$

The volume density of point defect scatterers is $n_{imp} = N_{imp}/L^3$, and the assumption of uncorrelated scattering sites as described in Equation 23.12 has been made. The scattering rate *increases* with carrier momentum $\hbar k$. To compute the resulting carrier mobility, we follow the recipe outlined in the last section to get

$$\tau_m^{pt}(\mathbf{k}) = \frac{\pi\hbar^3}{n_{imp}m^\star W_0^2\Omega_0^2}\left(\frac{1}{k}\right) = \underbrace{\frac{\pi\hbar^3}{n_{imp}m^\star W_0^2\Omega_0^2}\cdot\frac{\hbar}{\sqrt{2m^\star k_b T}}}_{\tau_0}\cdot\underbrace{\left(\frac{E_{\mathbf{k}}}{k_b T}\right)^{-\frac{1}{2}}}_{\eta^r}$$

$$(d=3, r=-\frac{1}{2}) \implies \mu_{3d}^{pt} = \frac{q\tau_0}{m^\star}\cdot\underbrace{\frac{\Gamma(\frac{3}{2}-\frac{1}{2}+1)}{\Gamma(\frac{3}{2}+1)}}_{\frac{4}{3\sqrt{\pi}}}\cdot\underbrace{\frac{F_{\frac{3}{2}-\frac{1}{2}-1}(\eta)}{F_{\frac{3}{2}-1}(\eta)}}_{\frac{\ln(1+e^\eta)}{F_{\frac{1}{2}}(\eta)}}$$

$$\implies \boxed{\mu_{3d}^{pt} = \frac{2\sqrt{2\pi}q\hbar^4}{3n_{imp}(m^\star)^{\frac{5}{2}}W_0^2\Omega_0^2(k_bT)^{\frac{1}{2}}}\cdot\frac{\ln(1+e^\eta)}{F_{\frac{1}{2}}(\eta)}}$$

$$\underbrace{\approx}_{\eta<<-1}\frac{2\sqrt{2\pi}q\hbar^4}{3n_{imp}(m^\star)^{\frac{5}{2}}W_0^2\Omega_0^2(k_bT)^{\frac{1}{2}}} \text{ and } \underbrace{\approx}_{\eta>>+1}\frac{\pi q\hbar^4}{\sqrt{2}n_{imp}(m^\star)^{\frac{5}{2}}W_0^2\Omega_0^2(E_F)^{\frac{1}{2}}}. \tag{23.38}$$

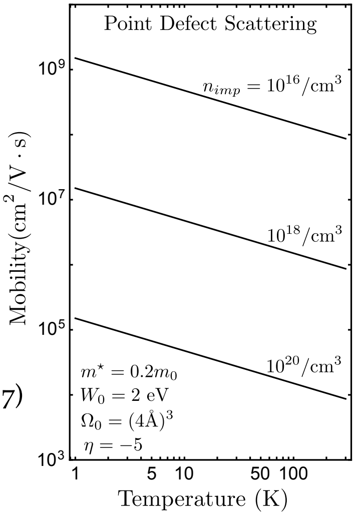


Fig. 23.20 Mobility limited by point defect scattering. The boxed Equation 23.38 is used for the calculation with the parameters indicated. For these parameters, if the point defect density is kept below $10^{18}/\text{cm}^3$, scattering by such defects will not affect the mobilities if values upto 1 million are allowed by other scattering mechanisms. The W_0 and Ω_0 must be matched to the particular defects of interest such as neutral vacancies, interstitials, or isovalent substitutional impurity atoms.

The final analytical results directly yield the carrier mobility in terms of known parameters. Fig. 23.20 shows example values.

2D carriers: If the mobile carriers are confined by barriers to move in two dimensions, such as in quantum wells, they will only scatter from those point defects that spatially overlap with the location of the carriers. This is a result of the short-range nature of point-defect scatterers. To compute the matrix element, we write the location of the scatterer as $\mathbf{r}_i = (\vec{\rho}_i, z_i)$ where $\vec{\rho}$ is a vector in the (x,y) plane in which the carriers are confined. Writing the envelope wavefunction of the 2D carriers as $\psi(\vec{\rho}, z) = \chi(z) \cdot [\frac{1}{\sqrt{L^2}} e^{i\mathbf{k}\cdot\vec{\rho}}]$, the squared matrix element is

$$\begin{aligned}
W_i(\mathbf{r}) &= W_0\Omega_0\delta(\mathbf{r}-\mathbf{r}_i) = W_0\Omega_0\delta(\vec{\rho}-\vec{\rho}_i)\delta(z-z_i) \\
W_i(q) &= W_0\Omega_0 \int d^3r\psi_f\delta(\vec{\rho}-\vec{\rho}_i)\delta(z-z_i)\psi_i \\
&= W_0\Omega_0 I_{\mathbf{k},\mathbf{k}'} \underbrace{\int dz\chi_f^\star(z)\chi_i(z)\delta(z-z_i)}_{\chi_f^\star(z_i)\chi_i(z_i)} \underbrace{\int d^2r\frac{e^{i\mathbf{q}_{||}\vec{\rho}}}{L^2}\delta(\vec{\rho}-\vec{\rho}_i)}_{e^{\mathbf{q}_{||}\cdot\vec{\rho}_i}/L^2} \\
|W_{tot.}(q)|^2 &= \frac{W_0^2\Omega_0^2}{L^4}(e^{i\mathbf{q}_{||}\cdot\vec{\rho}_1}\chi^2(z_1) + e^{i\mathbf{q}_{||}\cdot\vec{\rho}_2}\chi^2(z_2) + ...) \times \\
&\qquad (e^{-i\mathbf{q}_{||}\cdot\vec{\rho}_1}\chi^2(z_1) + e^{-i\mathbf{q}_{||}\cdot\vec{\rho}_2}\chi^2(z_2) + ...) \\
= \frac{W_0^2\Omega_0^2}{L^4}&[\underbrace{\chi^4(z_1) + \chi^4(z_2) + ...\chi^4(z_{N_{imp}})}_{L^2\cdot\int dz\cdot n_{imp}(z)\cdot\chi^4(z)} + \underbrace{\sum_{m,n} e^{i\phi_{mn}}\chi_m^2(z)\chi_n^2(z)}_{=0\text{, sum over random phases }\phi_{mn}}] \\
|W_{tot.}(q)|^2 &= \frac{W_0^2\Omega_0^2}{L^2}\underbrace{\int dz\cdot n_{imp}(z)\cdot|\chi(z)|^4}_{n_{imp}^{eff.}}.
\end{aligned} \tag{23.39}$$

The 2D scattering rate is then obtained using Equation 23.21:

$$\frac{1}{\tau_m(\mathbf{k})} = \frac{W_0^2\Omega_0^2 n_{imp}^{eff.} m^\star}{2\pi(\hbar k)^3} \underbrace{\int_0^{2k} dq \frac{q^2}{\sqrt{1-(\frac{q}{2k})^2}}}_{2\pi k^3} = \frac{W_0^2\Omega_0^2 m^\star n_{imp}^{eff.}}{\hbar^3}$$

$$\implies \boxed{\mu_{2d}^{pt} = \frac{q\hbar^3}{W_0^2\Omega_0^2(m^\star)^2 n_{imp}^{eff}},} \tag{23.40}$$

which is an exact expression for carrier mobility with $(d = 2, r = 0)$. If the point defects are distributed uniformly in 3D with density n_{3d} in cm^{-3} units, then the factor $n_{imp}^{eff} = \frac{n_{3d}}{L_{eff}}$ in units of cm^{-4}, and $1/L_{eff} = \int dz \cdot |\chi(z)|^4$ only depends on the electron envelope function in the direction of confinement. Though the general case must be evaluated numerically, for lowest subband occupation with the Fang–Howard–Stern envelope function $L_{eff} = \frac{16}{3b}$, and for an infinitely deep square well of thickness L_z, $L_{eff} = \frac{2L_z}{3}$ as listed in Table 23.2.

Our treatment of point defect scattering assumes that the scattering event does not change the "internal coordinates" of the point scatterer itself. It is possible for the assumption of elastic scattering to be violated at high electron energies, or in magnetic semiconductors. For example, if an electron bound in the point scattering site is excited to a higher energy state upon collision with a mobile electron, the mobile electron loses this energy[13].

[13] This form of impact ionization can produce luminescence when the excited, but localized electron relaxes back to its ground state by emitting a photon. Similar scattering can also change the bonding configuration, or launch phonon waves. We do not consider such scattering here that typically requires the mobile electron to have a very high kinetic energy.

23.4 Coulomb impurity scattering

3D Carriers: Unlike the charge-neutral point-defect scattering potential, that due to ions such as donors or acceptors at a distance r are long-range Coulomb potentials that fall off as $1/r$ if the potential is unscreened, and as $e^{-q_s r}/r$ when screened, as we derived in Equation 23.7. Using Equation 23.19 this leads to a 3D scattering rate

$$W(r) = \frac{e^2}{4\pi\epsilon_s r}e^{-q_s r} \implies W(q) = \frac{e^2}{\epsilon_s(q^2+q_s^2)L^3}$$

$$\implies \frac{1}{\tau_m^{coul}(\mathbf{k})} = n_{imp}\frac{m^\star e^4}{4\pi\epsilon_s^2(\hbar k)^3}\int_{q=0}^{2k} dq \cdot \frac{q^3}{(q^2+q_s^2)^2}$$

$$\implies \boxed{\frac{1}{\tau_m^{coul}(\mathbf{k})} = n_{imp}\frac{m^\star e^4}{8\pi\epsilon_s^2(\hbar k)^3}\cdot\left(\ln(1+\frac{4k^2}{q_s^2}) - \frac{4k^2}{4k^2+q_s^2}\right).} \tag{23.41}$$

Labeling the term in the large brackets as $C(k)$, we note that its dependence on k is very weak as may be verified by plotting for typical values of the screening wavevectors. Since $C(k)$ is almost constant (say $C(k_0)$) with electron kinetic energy, the energy dependence is dominated by the $(\hbar k)^3$ term in the denominator via $E_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m^\star}$, which therefore is a power-law dependence on energy with $r = +\frac{3}{2}$. Using this fact, the electron mobility is obtained as

$$\boxed{\mu_{3d}^{coul} = \frac{128\sqrt{2\pi}\epsilon_s^2(k_b T)^{\frac{3}{2}}}{n_{imp}e^3\sqrt{m^\star}C(k_0)}\cdot\frac{F_2(\eta)}{F_{\frac{1}{2}}(\eta)}}$$

$$\implies \mu_{3d}^{coul} \underbrace{\approx}_{\eta<<-1} \frac{128\sqrt{2\pi}\epsilon_s^2(k_b T)^{\frac{3}{2}}}{n_{imp}e^3\sqrt{m^\star}C(k_0)}, \text{ or } \mu_{3d}^{coul} \underbrace{\approx}_{\eta>>+1} \frac{96\sqrt{2\pi}\epsilon_s^2(E_F)^{\frac{3}{2}}}{n_{imp}e^3\sqrt{m^\star}C(k_0)}. \tag{23.42}$$

For the non-degenerate limit, the classic $\mu_{3d}^{coul} \sim T^{\frac{3}{2}}$ dependence is thus obtained for 3D carriers of low density. The $1/n_{imp}$ dependence on the impurity density implies lower mobilities for higher doping densities, if the carriers are thermally ionized from the donors or acceptors. On the other hand, for degenerately doped carriers, the temperature dependence vanishes, and the mobility changes as $E_F^{\frac{3}{2}}$. Since $E_F = \frac{\hbar^2 k_F^2}{2m^\star}$ and $k_F = (3\pi^2 n_{3d})^{\frac{1}{3}}$, the mobility changes roughly as $\mu_{3d}^{coul} \sim \frac{n_{3d}}{n_{imp}}$. This is a characteristic feature of the carrier mobility in heavily doped semiconductors: it has a very weak temperature dependence all the way down to cryogenic temperatures. Note also the strong dependence on the relative dielectric constant: semiconductors with high ϵ_s are less affected by Coulomb impurity scattering. Fig. 23.22 shows an example of mobility limited by Coulomb impurity scattering.

Long-range Coulomb potential favors small-angle scattering, as may be concluded by examining the scattering potential in Equation 23.41

Fig. 23.21 Esther Conwell formulated the early understanding of impurity scattering and high-field electron transport phenomena in semiconductors. The approach of Conwell and Weisskopf appeared before that by Brooks and Herring, which is discussed here.

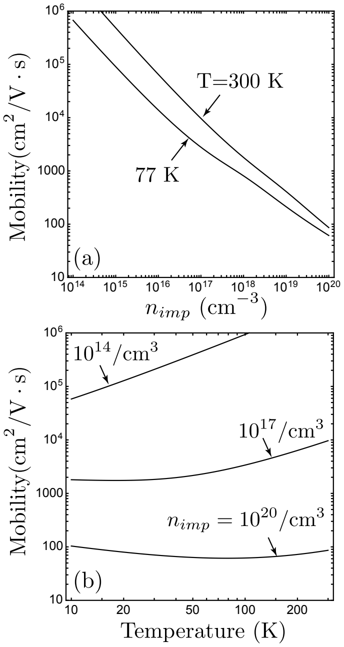


Fig. 23.22 Electron mobility limited by Coulomb impurity scattering in GaN (a) as a function of impurity density and (b) as a function of temperature. The free carrier concentration is assumed equal to the impurity concentration. The $\mu \sim T^{\frac{3}{2}}$ dependence may only be observed in extremely pure samples of very low carrier concentrations in the non-degenerate regime.

with $q = 2k\sin\frac{\theta}{2}$. Fig. 23.23 highlights this effect: the angular deviation for fast-moving electrons with large k is small, whereas slow-moving electrons undergo much larger angular deviation. This property makes the momentum relaxation rate (which depends on the angle of scattering) different from the quantum relaxation rate (which has no dependence on the scattering angle).

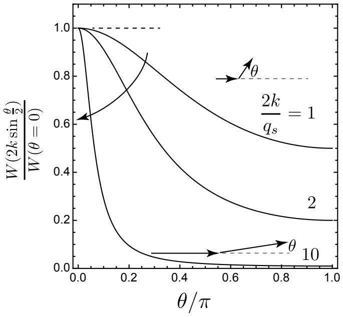


Fig. 23.23 Coulomb scattering for high energy electrons occurs by small angles. The Coulomb scattering rates were first evaluated by Conwell (Fig. 23.21) and Weiskoff, and by Brooks and Herring. The scattering rate models go by their names.

2D carriers: For carriers quantum confined in 2D the long-range Coulomb potential causes scattering even if the scatterer is remote, but with a significantly reduced strength. This type of remote impurity scattering occurs in modulation-doped heterostructures. Consider a single Coulomb scatterer located a distance d_0 as depicted in Fig. 23.14 from a confined carrier distribution. Using $\mathbf{q}_{\|} = \mathbf{k}_{\|} - \mathbf{k}'_{\|}$ and ϕ as the angle between 2D vectors $\mathbf{q}_{\|}$ and $\vec{\rho}$ (see Fig. 23.12), the scattering potential and the corresponding matrix element for the 2D carrier is

$$W_{unscr.}(\rho, d_0) = \frac{e^2}{4\pi\epsilon_s\sqrt{\rho^2 + (d_0+z)^2}} \implies \langle\psi_f|W_{unscr.}(\rho, d_0)|\psi_i\rangle =$$

$$\int d^3r[u^\star_{cf}(r)\chi^\star_f(z)\frac{e^{-i\mathbf{k}'_{||}\cdot\vec{\rho}}}{L^2}]\cdot[\frac{q^2}{4\pi\epsilon_s\sqrt{\rho^2+(d_0+z)^2}}]\cdot[u_{ci}(r)\chi_i(z)\frac{e^{+i\mathbf{k}_{||}\cdot\vec{\rho}}}{L^2}]$$

$$= \frac{e^2}{4\pi\epsilon_s L^2}\int_{z=0}^{\infty} dz\chi^\star_f(z)\chi_i(z)\underbrace{\int_{\rho=0}^{\infty}\int_{\phi=0}^{2\pi}\frac{\rho d\rho d\phi e^{iq_{||}\rho\cos\phi}}{\sqrt{\rho^2+(d_0+z)^2}}}_{\frac{2\pi}{q_{||}}e^{-q_{||}(d_0+z)}}$$

$$\implies W_{unscr.}(q_{||}, d_0) = \frac{e^2 e^{-q_{||}d_0}}{2\epsilon_s L^2 q_{||}}\underbrace{\int dz\chi^\star_f(z)\chi_i(z)e^{-q_{||}z}}_{P_0(q_{||})}$$

$$\implies W_{scr}(q_{||}, d_0) = \frac{W_{scr}(q_{||}, d_0)}{1+\frac{q_s}{q_{||}}G(q_{||})} = \frac{e^2 e^{-q_{||}d_0}P_0(q_{||})}{2\epsilon_s L^2[q_{||}+q_s P_1(q_{||})]}, \tag{23.43}$$

where the last equation is the 2D Coulomb matrix element for the *screened* potential, with the form factors $P_0(...)$ and $P_1(...)$ described in Equation 23.16. Thus, the remote impurity scattering rate is

$$\frac{1}{\tau_m(k_{\|})} = \frac{N_{imp}L^2 m^\star}{2\pi(\hbar k_{\|})^3}\int_0^{2k_{\|}} dq_{\|}\frac{q_{\|}^2|W_{scr}(q_{\|}, d_0)|^2}{\sqrt{1-(\frac{q_{\|}}{2k_{\|}})^2}}$$

$$= \underbrace{\frac{N_{imp}}{L^2}}_{n_{imp}}\cdot\frac{m^\star e^4}{4\pi\epsilon_s^2\hbar^3 k_{\|}^2}\cdot\underbrace{\int_0^1 du\frac{u^2 e^{-2k_{\|}d_0 u}P_0^2(2k_{\|}u)}{\sqrt{1-u^2}(u+\frac{q_s}{2k_{\|}}P_1(2k_{\|}u))^2}}_{I(k_{\|}, n_{2d}, d_0, q_s)}, \tag{23.44}$$

where $u = q_{\|}/2k_{\|}$ is used. The integral $I(k_{\|}, n_{2d}, d_0, q_s)$ has no simple analytical form, though some limits for a delta-function confined well were outlined in Table 23.4. Writing the scattering rates in terms of this integral and $E_{\mathbf{k}} = \frac{\hbar^2 \mathbf{k}_\parallel}{2m^\star}$, we obtain the 2D carrier mobility

$$\frac{1}{\tau_m(k_\parallel)} = \frac{n_{imp} m^\star e^4 I(k_\parallel, n_{2d}, d_0, q_s)}{4\pi\epsilon_s^2 \hbar^3 k_\parallel^2} = \frac{n_{imp} e^4 I(E_{\mathbf{k}}, n_{2d}, d_0, q_s)}{8\pi\epsilon_s^2 \hbar E_{\mathbf{k}}},$$

$$\implies \mu_{2d}^{coul} = \frac{e}{m^\star k_b T} \frac{\int_{E_{\mathbf{k}}=0}^{\infty} dE_{\mathbf{k}} E_k f_{\mathbf{k}}^0 (1 - f_{\mathbf{k}}^0) \tau_m(E_{\mathbf{k}})}{\int_{E_{\mathbf{k}}=0}^{\infty} dE_{\mathbf{k}} f_{\mathbf{k}}^0} \approx \frac{e\tau_m(E_F)}{m^\star}.$$

$$\mu_{2d}^{coul} \approx \frac{8\pi^2 \epsilon_s^2 \hbar^3 n_s}{e^3 (m^\star)^2 n_{imp}} \cdot \frac{1}{I(k_F, n_{2d}, d_0, q_s)} \approx (2k_F d_0)^3 \frac{e}{\pi\hbar n_{imp}}, \tag{23.45}$$

where the last approximate form is valid for $d_0 >> 1/(2k_F)$. This limit is physically realized when the scattering ion is more distant from a mobile 2D electron (or hole) than the next mobile electron in the 2D plane. The reduction of the ionized impurity scattering rate because of modulation doping is very significant, and is responsible for some of the highest mobilities achieved in semiconductors. Fig. 23.24 shows the mobility achievable with remote impurity scattering. It highlights the increase in mobility by spatial separation of the sheet of dopants from the 2D carriers. This increase asymptotically approaches $\sim d_0^3$ for large d_0, as expected from Equation 23.45.

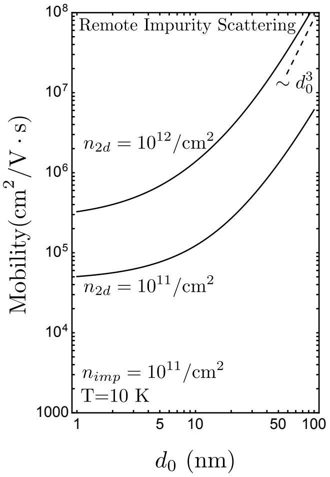


Fig. 23.24 Remote Coulomb scattering for AlGaAs/GaAs 2D electron gases, with the distance d_0 defined in Fig. 23.14. The highest mobilities achieved in such 2DEGs exceeds 10s of million at low temperatures, and therefore even the remote scattering must be carefully engineered.

If the Coulomb impurity scatterers are distributed as background impurities in 3D as $n_{imp}^{3d}(z)$, the net scattering rate for 2D confined carriers is obtained by integrating the scattering matrix element over the distribution. For example, if the dopants are distributed uniformly in 3D, the net scattering rate becomes

$$\begin{aligned}\frac{1}{\tau_m(k_\parallel)} &= \frac{m^\star e^4}{4\pi\epsilon_s^2 \hbar^3 k_\parallel^2} \int_0^1 \frac{u^2 P_0^2(2k_\parallel u) du}{\sqrt{1-u^2}(u + \frac{q_s}{2k_\parallel} P_1(2k_\parallel u))^2} \cdot \int n_{imp}^{3d}(z) e^{-2k_\parallel z u} dz \\ &= \frac{m^\star n_{imp}^{3d} e^4}{4\pi\epsilon_s^2 \hbar^3 k_\parallel^2} \int_0^1 du \frac{u^2 P_0^2(2k_\parallel u)}{\sqrt{1-u^2}(u + \frac{q_s}{2k_\parallel} P_1(2k_\parallel u))^2} \cdot \underbrace{\int_0^\infty dz \cdot e^{-2k_\parallel z u}}_{=\frac{1}{2k_\parallel u}} \\ &= \frac{m^\star n_{imp}^{3d} e^4}{8\pi\epsilon_s^2 \hbar^3 k_\parallel^3} \int_0^1 du \frac{u P_0^2(2k_\parallel u)}{\sqrt{1-u^2}(u + \frac{q_s}{2k_\parallel} P_1(2k_\parallel u))^2},\end{aligned} \tag{23.46}$$

which then directly yields the carrier mobility. It is important to keep the background Coulomb impurity level n_{imp}^{3d} well below say $10^{14}/\text{cm}^3$ to suppress this form of scattering below other scattering mechanisms. Just like for 3D bulk carriers, both remote and background Coulomb impurities favor small-angle scattering heavily, and therefore also have characteristically large ratios of momentum to quantum to scattering times, i.e., $\frac{\tau_m}{\tau_q} > 1$.

23.5 Dipole scattering

3D carriers: Another form of Coulomb scattering occurs in the presence of charge dipoles in semiconductors when equal and opposite charge defects exist in pairs in close proximity. This may be, for example, due to donor-acceptor pairs in compensated semiconductors, or due to strongly localized excitons. Dipole scattering also occurs in semiconductor alloys that possess spontaneous or piezoelectric polarization, since each unit cell is then modeled to contain a charge dipole characterized by charges $+e$ and $-e$ separated by a distance d. If the semiconductor is an alloy, the alloy composition determines the strength of the dipole, making the dipole vector $\mathbf{p} = (e \cdot d)\hat{\mathbf{n}}$ spatially vary in every unit cell depending on the local alloy composition. Here $\hat{\mathbf{n}}$ is the unit vector in the direction of the dipole. The direction dependence of dipole scattering is distinct from Coulomb scattering from single point charges which are isotropic. If the semiconductor is ferroelectric (or polycrystalline), the direction of $\mathbf{p}$ can change in space based on the domain orientation (or local crystal orientation).

An electron at a point (r, β) from the center of the dipole will experience a potential that is the sum of the positive and negative charges. With screening, this leads to the momentum scattering rate using the established recipe:

$$\begin{aligned}
W(r) &= \frac{e^2}{4\pi\epsilon_s}\left[\frac{1}{r - \frac{d}{2}\cos\beta} - \frac{1}{r + \frac{d}{2}\cos\beta}\right] \approx \frac{e^2}{4\pi\epsilon_s}\frac{d\cos\beta}{r^2} \\
W(q) &= \frac{e^2 d\cos\beta}{4\pi\epsilon_s}\int \frac{d^3r}{L^3} e^{i\mathbf{q}\cdot\mathbf{r}} \cdot \frac{1}{r^2} = \frac{\pi e^2 d\cos\beta}{2\epsilon_s L^3 q} \\
\implies W_{scr}(q) &= \frac{\pi e^2 d\cos\beta}{2\epsilon_s L^3 (q + q_s)} \text{ where } q_s = \frac{e^2 d\langle\cos\beta\rangle N_c^{3d} F_{-\frac{1}{2}}(\eta)}{16\pi^2\epsilon_s k_b T} \\
\frac{1}{\tau_m^{dip}(\mathbf{k})} &= \frac{3\pi e^4 m^\star d^2 \langle\cos^2\beta\rangle n_{imp}}{8\epsilon_s^2\hbar^3 k} \times \\
&\Big(1 + \frac{1}{2}(\frac{q_s}{k})^2 \ln[1 + \frac{2}{\frac{q_s}{k}}] - \frac{q_s}{k} - \frac{4}{3(2 + \frac{q_s}{k})}\Big),
\end{aligned} \tag{23.47}$$

from where the dipole-scattering limited electron mobility can be computed. Note that if the dipoles are arranged randomly, or if the carriers are allowed to move in all directions, the $\langle\cos^2\beta\rangle$ can be suitably integrated to yield constant factors of $1/3$ for 3D and $1/2$ for 2D.

2D carriers: Mobility limited by dipole scattering for 2D carriers for both remote and background dipole densities is calculated by following the procedure similar to the corresponding discussion of Section 23.4. Because the dipole potential decays faster than the single charge Coulomb potential, this scattering is important if their density is high, e.g. for dipoles of spatially varying strength at near atomic densities in polar semiconductor alloys or ferroelectric semiconductors.

23.6 Dislocation scattering

Dislocations are topological 1D defects in 3D semiconductor crystals. They form to relax strain in the crystal, and are characterized by a linear array of broken chemical bonds, surrounded by a strain field. As shown in Fig. 23.25, the dangling bonds can capture free–electrons from the conduction band (or holes from the valence band), and form a line charge of density $\lambda_L = ef/c_0$, where e is the electron charge, c_0 is the lattice constant along the dislocation axis, and f is the fractional occupancy of such sites.

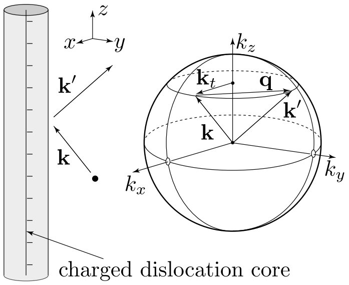


Fig. 23.25 Scattering sphere for dislocation scattering. For a dislocation aligned along the z-direction, the electron momentum represented by its wavevector can only change in the $x-y$ direction, which means the scattered wavevector $\mathbf{k}'$ lies on a latitude of the scattering sphere.

3D carriers : Electrons moving in 3D around such charged dislocations will scatter due to the screened Coulomb potential, which by solving the screened Poisson equation in cylindrical geometry is found to be of the Bessel $K_0(x)$ function of zeroth order, giving

$$W_{scr}(\rho) = \frac{e\lambda_L}{2\pi\epsilon_s} K_0(\frac{\rho}{L_D}) \implies W_{scr}(q) = \frac{e\lambda_L}{2\pi\epsilon_s} \int \frac{\rho d\rho d\phi}{L_x L_y} e^{i\mathbf{q}\cdot\rho} K_0(\frac{\rho}{L_D})$$
$$= \frac{e\lambda_L}{2\pi\epsilon_s L_x L_y} \int_{\rho=0}^{\infty} d\rho \cdot \rho K_0(\frac{\rho}{L_D}) \underbrace{\int_{\phi=0}^{2\pi} d\phi e^{iq\rho\cos\phi}}_{2\pi J_0(q\rho)}$$
$$= \frac{e\lambda_L}{\epsilon_s L_x L_y} \underbrace{\int_{\rho=0}^{\infty} d\rho \cdot \rho K_0(\frac{\rho}{L_D}) J_0(q\rho)}_{\frac{1}{q^2+q_D^2}} \implies W_{scr}(q) = \frac{e\lambda_L}{\epsilon_s L_x L_y} \cdot \frac{1}{q^2+q_D^2}, \tag{23.48}$$

where $L_D = 1/q_D$ is the 3D screening length. To find the momentum scattering rate of state $\mathbf{k}$ the transition rates of all final states $\mathbf{k}'$ must be summed. Unlike Coulomb scattering from 3D point charges, that due to a uniform linear charge along the z-axis restricts the number of final $\mathbf{k}'$ states to those that have the same k_z as the initial state, since the momentum along the dislocation axis cannot change. Fig. 23.25 shows that we must then sum over constant transverse wavevector k_t "latitude circles" instead of the entire surface of the sphere to get:

$$\frac{1}{\tau_m^{disl.}(k_t)} = \frac{N_{disl.} L_x L_y m^\star}{2\pi(\hbar k)^3} \int_{q=0}^{2k_t} dq \frac{q^2 |W_{scr}(q)|^2}{\sqrt{1-(\frac{q}{2k_t})^2}}$$
$$= \frac{N_{disl.} L_x L_y m^\star}{2\pi(\hbar k_t)^3} \frac{e^2\lambda_L^2}{\epsilon_s^2 (L_x L_y)^2} \underbrace{\int_{q=0}^{2k_t} dq \frac{q^2}{(q^2+q_D^2)^2 \sqrt{1-(\frac{q}{2k_t})^2}}}_{\frac{2\pi k_t^3}{q_D (4k_t^2+q_D^2)^{\frac{3}{2}}}}$$
$$\implies \boxed{\frac{1}{\tau_m^{disl.}(k_t)} = \frac{n_{disl.} m^\star e^2 \lambda_L^2}{\epsilon_s^2 \hbar^3 q_D^4 (1+\frac{4k_t^2}{q_D^2})^{\frac{3}{2}}}}. \tag{23.49}$$

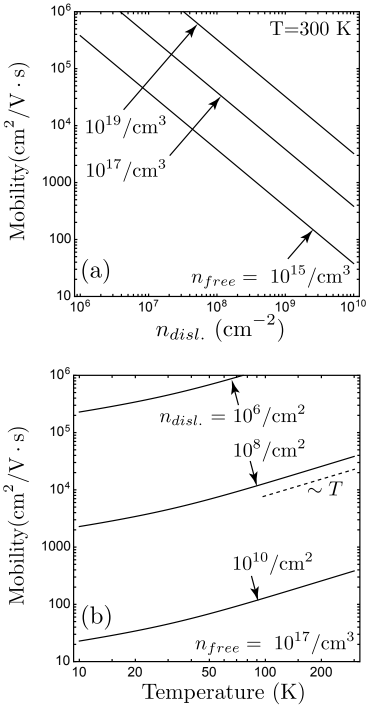


Fig. 23.26 Electron mobility in 3D bulk GaN limited by dislocation scattering (a) as a function of dislocation density for various free carrier concentrations and (b) as a function of temperature for various dislocation densities. Note the $\mu \sim T$ dependence in (b) at high temperatures derived in Equation 23.50.

The carrier mobility can be obtained numerically since the momentum relaxation time is not a polynomial function of energy as required to apply Equation 23.35. For non-degenerate conditions when $2k_t >> q_D$, which is achieved at temperatures at which the thermal energy is larger than the plasmon energy $k_b T \geq \hbar\omega_{pl}$ (see Table 23.1), the polynomial form is obtained, resulting in an approximate analytical form of mobility limited by dislocation scattering (see Fig. 23.26):

$$\mu_{3d}^{disl.} \approx \frac{512\sqrt{\pi}}{3\sqrt{3}} \cdot \frac{q^2 c_0^2 \epsilon_s^{\frac{3}{2}} n_{3d}^{\frac{1}{2}} k_b T}{n_{disl.} (m^\star)^{\frac{1}{2}} e^4 f^2} \sim \frac{n_{3d}^{\frac{1}{2}} T}{n_{disl.}}. \tag{23.50}$$

2D carriers: The unscreened electric field in the 2D plane of carriers a distance ρ from the dislocation core (see Fig. 23.12) gives:

$$E(\rho) = \frac{\lambda_L}{2\pi\epsilon_s \rho} \implies W_{uns.}(\rho) = -q\int d\rho E(\rho) = -\frac{q\lambda_L}{2\pi\epsilon_s}\ln(\frac{\rho}{L})$$

$$W_{unsc}(q_{||}) = -\frac{q\lambda_L}{2\pi\epsilon_s L^2}\int_{\rho=0}^{\infty} d\rho \cdot \rho \ln(\frac{\rho}{L}) \underbrace{\int_{\phi=0}^{2\pi} d\phi e^{i q_{||} \rho \cos(\phi)}}_{2\pi J_0(q_{||}\rho)}$$

$$= -\frac{q\lambda_L}{\epsilon_s L^2} \underbrace{\int_0^{\infty} d\rho \cdot \rho \ln(\rho) \cdot J_0(q_{||}\rho)}_{-1/q_{||}^2}$$

$$\implies W_{uns.}(q_{||}) = \frac{q\lambda_L}{\epsilon_s L^2 q_{||}^2} \implies W_{scr.}(q_{||}) = \frac{q\lambda_L P_0(q_{||})}{\epsilon_s L^2 q_{||}^2 (1 + \frac{q_s}{q_{||}} P_1(q_{||}))}. \tag{23.51}$$

This screened matrix element gives the scattering rate and mobility:

$$\frac{1}{\tau_m^{disl.}(k_{||})} = \frac{N_{disl.} L^2 m^\star}{2\pi(\hbar k_{||})^3} \int_0^{2k_{||}} dq_{||} \frac{q_{||}^2 |W_{scr.}(q_{||})|^2}{\sqrt{1 - (\frac{q_s}{2k_{||}})^2}}$$

$$\frac{1}{\tau_m^{disl.}(k_{||})} = \underbrace{\frac{N_{disl.}}{L^2}}_{n_{disl.}} \cdot \frac{m^\star q^2 \lambda_L^2}{\epsilon_s^2 \hbar^3} \cdot \frac{1}{4\pi k_{||}^4} \cdot \underbrace{\int_0^1 du \frac{P_0^2(2k_{||}u)}{\sqrt{1-u^2}(u + \frac{q_s}{2k_{||}} P_1(2k_{||}u))^2}}_{I_{disl.}^{ch}(k_{||})}$$

$$\frac{1}{\tau_m^{disl.}(k_{||})} = n_{disl.} \cdot \frac{m^\star q^2 \lambda_L^2}{\epsilon_s^2 \hbar^3} \cdot \frac{I_{disl.}^{ch}(k_{||})}{4\pi k_{||}^4} = n_{disl.} \cdot \frac{q^2 \lambda_L^2 \hbar}{16\pi\epsilon_s^2 m^\star E_{\mathbf{k}}^2} \cdot I_{disl.}^{ch}(E_{\mathbf{k}})$$

$$\mu_{2d}^{disl} \approx \frac{q\tau_m(E_F)}{m^\star} \approx \frac{16\pi\epsilon_s^2 m^\star E_F^2}{q\lambda_L^2 \hbar n_{disl.} I_{disl.}^{ch}(E_F)} = \frac{4\pi\epsilon_s^2 \hbar^3 k_F^4}{q\lambda_L^2 (m^\star)^2 n_{disl.} I_{disl.}^{ch}(k_F)}. \tag{23.52}$$

Fig. 23.27 shows the effect of charged dislocation scattering on the mobility of 2D electron gases[14].

Dislocations also introduce strain fields around them that decay radially from the core as $\sim \sin\theta/\rho$ where θ is the angle between the

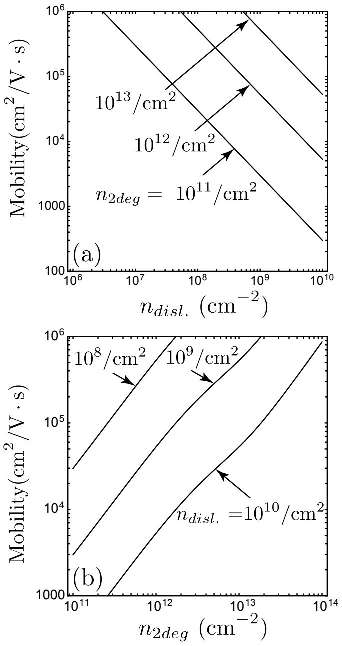


Fig. 23.27 Electron mobility in AlGaN/GaN 2DEGs limited by dislocation scattering evaluated using a Fang–Howard–Stern variational wavefunction and $f = 1$. Since room-temperature electron mobilities in such heterostructures are limited to $\sim$ 2000 cm^2/V·s by phonons, they are nearly unaffected by dislocation densities of $n_{disl.} \sim 10^8/\text{cm}^2$. The formulation is for dislocations that run perpendicular to the 2D plane: this is easily modified for dislocations that cross the plane at angles.

[14] $I_{disl.}^{ch}(...)$ in Equation 23.52 requires numerical integration. Approximate cases may be found using Table 23.4. The numerical dependence on 2D carrier density of $\sim n_{2deg}^{1.34}$ in Fig. 23.27 can be approximated by $\sim n_{2deg}^{3/2}$ for an electron gas confined in a perfect 2D plane.

Burgers vector $\mathbf{b}_e$ and the radial vector $\vec{\rho}$. Such strain-fields around dislocations lead to a scattering rate and mobility[15]

$$\Delta E_c = D_c \text{Trace}(\epsilon) \implies W(\rho,\theta) = \Delta E_c = -\frac{b_e D_c}{2\pi}\left(\frac{1-2\gamma}{1-\gamma}\right)\frac{\sin\theta}{\rho}$$

$$\implies \frac{1}{\tau_m^{disl.}(k_\|)} = \frac{n_{disl.} m^\star b_e^2 D_c^2}{2\pi k_\|^2 \hbar^3}\left(\frac{1-2\gamma}{1-\gamma}\right)^2 \int_0^1 du \frac{u^2 P_0^2(u)}{\sqrt{1-u^2}[u+\frac{q_{TF}}{2k_\|}P_1(u)]^2}$$

$$\implies \mu_{2d}^{disl.,str} \approx \frac{2e\hbar^3 \pi k_F^2}{n_{disl.}(m^\star)^2 b_e^2 D_c^2}\left(\frac{1-\gamma}{1-2\gamma}\right)^2 \frac{1}{I_{disl.}^{str.}(n_{2d})}. \tag{23.53}$$

23.7 Alloy disorder scattering

Consider semiconductor alloys of the form $A_x B_{1-x} C$ for compound semiconductors[16] such as $\text{In}_x\text{Ga}_{1-x}\text{As}$ or $\text{Al}_x\text{Ga}_{1-x}\text{N}$. Alloy disorder caused by spatially random locations r_a of A atoms and r_b of B atoms (the C sublattice is not random) produces a microscopic potential

$$V(r) = \sum_{r_i=r_a} V_a(r-r_i) + \sum_{r_i=r_b} V_b(r-r_i) + \sum_{r_i=r_c} V_c(r-r_i). \tag{23.54}$$

Assuming the mole fraction $x = N_A/(N_A+N_B)$ where N_A, N_B, N_C are the number of atoms in the stoichiometric crystal with $N_A + N_B = N_C$, the spatially averaged potential which does not scatter electrons is

$$V_{av}(r) = \sum_{r_i=r_a,r_b} \frac{N_A V_a(r-r_i) + N_B V_b(r-r_i)}{N_A+N_B} + \sum_{r_i=r_c} V_c(r-r_i)$$

$$= \sum_{r_i=r_a,r_b} [xV_a(r-r_i) + (1-x)V_b(r-r_i)] + \sum_{r_i=r_c} V_c(r-r_i). \tag{23.55}$$

The difference between the microscopically disordered potential and the spatially average potential is therefore the scattering potential for mobile carriers in the alloy:

$$W(r) = V(r) - V_{av}(r) =$$

$$\sum_{r_i=r_a}(1-x)\underbrace{(V_a(r-r_i) - V_b(r-r_i))}_{V_0\Omega_0\delta(r-r_i)} + (-x)\sum_{r_i=r_b}(V_a(r-r_i) - V_b(r-r_i))$$

$$\implies W(r) = V_0\Omega_0[(1-x)\sum_{r_i=r_a}\delta(r-r_i) - x\sum_{r_i=r_b}\delta(r-r_i)], \tag{23.56}$$

where V_0 is the on-site potential difference if atom A is replaced by B, or vice versa, and Ω_0 is the effective volume that atom A (or atom B) occupies in the crystal. For example, for the cubic $\text{Al}_x\text{Ga}_{1-x}\text{As}$ crystal of lattice constant a_0 the scattering volume $\Omega_0 = a_0^3/8$, because in the primitive cell of volume of $a_0^3/4$, there are 2 atoms.

---

[15] The strain deforms the band edges by $\Delta E_c = D_c\text{Trace}(\epsilon)$ where ϵ is the strain tensor and D_c is the deformation potential. This perturbation scatters electrons, even when the dislocation core is not charged. The Burgers vector is typically equal to a lattice constant in magnitude. In Equation 23.53, γ is the dimensionless Poisson ratio of the crystal, and $I_{disl.}^{str.}(n_{2d})$ must be evaluated numerically except for very specific cases. Strictly one must add the dislocation charge and the strain potentials *before* evaluating the scattering rate; the presentation here neglects potential interference of these two scattering processes. Table A.2 in the Appendix lists deformation potentials (D_c) of semiconductors.

[16] For elemental semiconductors this would be of the form $A_x B_{1-x}$, such as $\text{Si}_x\text{Ge}_{1-x}$.

3D carriers: Assuming unity overlap integral of the periodic parts of the carrier wavefunctions, and V the total crystal volume, the matrix element and its square for the alloy scattering potential for 3D carriers is obtained from Equation 23.56 using $\int d^3 r e^{i\mathbf{q}\cdot\mathbf{r}}\delta(r-r_i)/V = e^{i\mathbf{q}\cdot\mathbf{r}_i}/V$:

$$\langle \mathbf{k}'|W(r)|\mathbf{k}\rangle = V_0 \frac{\Omega_0}{V}[(1-x)\sum_{\mathbf{r}_i=\mathbf{r}_a} e^{i\mathbf{q}\cdot\mathbf{r}_i} + (-x)\sum_{\mathbf{r}_i=\mathbf{r}_b} e^{i\mathbf{q}\cdot\mathbf{r}_i}]$$

$$\implies |\langle \mathbf{k}'|W(r)|\mathbf{k}\rangle|^2 = (\frac{V_0\Omega_0}{V})^2 \times$$

$$\Big[(1-x)^2 \underbrace{(\sum_{r_i=r_a} e^{i\mathbf{q}\cdot\mathbf{r}_i})(\sum_{r_i=r_a} e^{-i\mathbf{q}\cdot\mathbf{r}_i})}_{N_A = x\frac{V}{\Omega_0}} + x^2 \underbrace{(\sum_{r_i=r_b} e^{i\mathbf{q}\cdot\mathbf{r}_i})(\sum_{r_i=r_b} e^{-i\mathbf{q}\cdot\mathbf{r}_i})}_{N_B=(1-x)\frac{V}{\Omega_0}}$$

$$-x(1-x)\underbrace{(\sum_{r_i=r_a} e^{i\mathbf{q}\cdot\mathbf{r}_i})(\sum_{r_i=r_b} e^{-i\mathbf{q}\cdot\mathbf{r}_i})}_{=0} - x(1-x)\underbrace{(\sum_{r_i=r_b} e^{i\mathbf{q}\cdot\mathbf{r}_i})(\sum_{r_i=r_a} e^{-i\mathbf{q}\cdot\mathbf{r}_i})}_{=0}\Big]$$

$$\implies |\langle \mathbf{k}'|W(r)|\mathbf{k}\rangle|^2 = \frac{\Omega_0}{V}x(1-x)V_0^2. \tag{23.57}$$

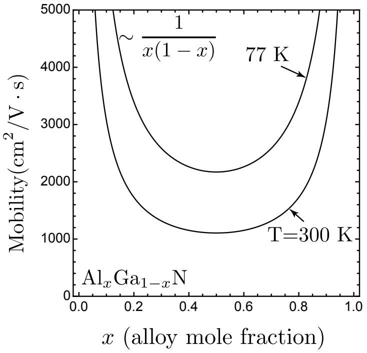


Fig. 23.28 Electron mobility in Al$_x$Ga$_{1-x}$N alloy limited by alloy disorder scattering. The scattering potential is $V_0 = 1.8$ eV, and $\Omega_0 = \frac{\sqrt{3}}{8}a^2 c$.

Note that this squared matrix element already includes *all* the scattering sites. The momentum scattering rate is

$$\frac{1}{\tau_m(\mathbf{k})} = \frac{2\pi}{\hbar}\sum_{\mathbf{k}'}|\langle \mathbf{k}'|W(r)|\mathbf{k}\rangle|^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}})(1-\cos\theta)$$

$$= \frac{2\pi}{\hbar}\frac{\Omega_0 x(1-x)V_0^2}{V}\int \frac{d^3k'}{\frac{(2\pi)^3}{V}}\delta(E_{\mathbf{k}'} - E_{\mathbf{k}})(1-\cos\theta)$$

$$= \frac{\Omega_0 x(1-x)m^\star V_0^2}{\pi\hbar^3}\underbrace{\int_{\theta=0}^{\pi} d\theta \sin\theta(1-\cos\theta)}_{=\int_{\theta=0}^{\pi} d\theta\sin\theta = 2}\underbrace{\int_{q=0}^{2k} dq \frac{q^2\delta((\frac{q}{2\sin(\frac{\theta}{2})})^2 - k^2)}{(2\sin(\frac{\theta}{2}))^3}}_{=k/2}$$

$$\implies \boxed{\frac{1}{\tau_m^{alloy}(k)} = \frac{\Omega_0 x(1-x)m^\star V_0^2 k}{\pi\hbar^3}}, \tag{23.58}$$

which is in the form $\tau(E) = \tau_0(\frac{E}{k_bT})^{-\frac{1}{2}}$. The electron mobility then[17] has the exact and approximate forms

$$\boxed{\mu_{3d}^{alloy} = \frac{\sqrt{\frac{8\pi}{9}}\cdot q\hbar^4 \cdot \ln(1+e^\eta)}{(m^\star)^{\frac{5}{2}}\Omega_0 x(1-x)V_0^2(k_bT)^{\frac{1}{2}}F_{\frac{1}{2}}(\eta)} \approx \frac{\sqrt{\frac{8\pi}{9}}q\hbar^4}{(m^\star)^{\frac{5}{2}}\Omega_0 x(1-x)V_0^2\sqrt{k_bT}}}. \tag{23.59}$$

The approximate form is for the non-degenerate case. Fig. 23.28 shows an example of alloy scattering limited mobility which goes as $\sim [x(1-x)]^{-1}$. As a result the mobility is the lowest near $x \sim 0.5$.

---

[17] Here we have used $d = 3$ and $r = -\frac{1}{2}$ to get $\mu_{3d}^{alloy} = \frac{e\tau_0}{m^\star}\frac{\Gamma(2)}{\Gamma(\frac{5}{2})}\frac{F_0(\eta)}{F_{\frac{1}{2}}(\eta)}$. Note the strong dependence of the alloy limited mobility on the effective mass and the scattering potential V_0, which is typically $\sim \Delta E_c$, the order of the band offset between the constituents of the alloy.

2D carriers: For carriers confined in the z-direction in quantum wells with envelope function $\chi(z)$, the well, the confining barrier, or both the well and the barrier can be alloys. We assume that the alloy composition x and the alloy-scattering potential V_0 are z-independent[18] . In this case, using Equation 23.57, the scattering rate is obtained as

$$\frac{1}{\tau_m^{alloy}} = \frac{m^\star \Omega_0 V_0^2 x(1-x)}{\hbar^3} \cdot \int_{z \in \text{alloy}} dz \cdot \chi^4(z), \tag{23.60}$$

which emphasizes that only if the alloy disorder potential overlaps spatially with the mobile carriers will it scatter them. If the 2DEG is in an alloy quantum well (e.g. InAlAs/InGaAs/InAlAs on InP substrates) with a Fang–Howard–Stern envelope function with variational parameter b , this immediately yields[19] the scattering rate and mobility

$$\frac{1}{\tau_m^{alloy}} = \frac{m^\star \Omega_0 V_0^2 x(1-x)}{\hbar^3} \cdot \frac{3b}{16} \text{ and } \boxed{\mu_{2d}^{alloy} = \frac{q\hbar^3}{(m^\star)^2 \Omega_0 V_0^2 x(1-x)} \cdot \frac{16}{3b}.} \tag{23.61}$$

On the other hand, if the mobile carriers are confined in a non-alloy quantum well, but there is some penetration into the barrier layer due to the finite potential barrier height (e.g. in AlGaAs/GaAs or AlGaN/GaN heterostructures), the envelope function $\chi(z)$ overlap with the alloy barrier must be evaluated numerically by solving Poisson and Schrödinger equations self-consistently. Simple analytical estimates may be obtained using a modified Fang–Howard–Stern envelope function to account for the finite penetration of the wavefunction into the alloy barrier[20]:

$$\chi(z) = Me^{\kappa_b z/2} \text{ for } z < 0 \text{ and } \chi(z) = N(z+z_0)e^{-bz/2} \text{ for } z \geq 0,$$

$$\text{where } \kappa_b = 2\sqrt{\frac{2m_b^\star \Delta E_c}{\hbar^2}},\; z_0 = \frac{2}{b + \kappa_b \frac{m_w^\star}{m_b^\star}},$$

$$N = \sqrt{\frac{b^3}{2}} \frac{1}{[1 + bz_0 + \frac{b^2 z_0^2}{2}(1 + \frac{b}{\kappa_b})]^{\frac{1}{2}}},\; M = Nz_0, \text{ and } P_b = \frac{N^2 z_0^2}{\kappa_b}. \tag{23.62}$$

This leads to the scattering rate and mobility

$$\frac{1}{\tau_m^{alloy}} = \frac{m^\star \Omega_0 V_0^2 x(1-x)}{\hbar^3} \cdot \frac{\kappa_b P_b^2}{2},\; \boxed{\mu_{2d}^{alloy} = \frac{q\hbar^3}{(m^\star)^2 \Omega_0 V_0^2 x(1-x)} \cdot \frac{2}{\kappa_b P_b^2}.} \tag{23.63}$$

Note that the final expression for the mobility is similar to Equation 23.61, except for the replacement of the effective length $16/3b$ by $2/(\kappa_b P_b^2)$. The penetration depth $2/\kappa_b$ in the barrier is where alloy scattering occurs, but P_b^2 in the denominator captures the fact that only the small fraction of mobile carriers that penetrate the alloy barrier are scattered, not the carriers that reside in the well (see Fig. 23.29).

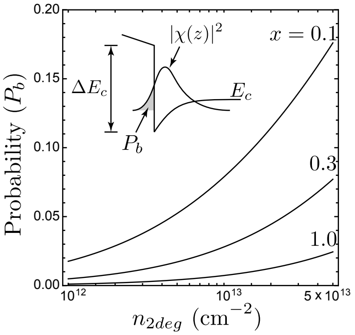


Fig. 23.29 Fraction of electrons P_b that penetrate the alloy barrier as a function of the 2D carrier density. The calculation uses Equations 23.62 for an Al$_x$Ga$_{1-x}$N barrier and a GaN well. The barrier penetration increases for small ΔE_c and high carrier densities.

[18] Note here x is *not* the spatial direction, but the mole fraction of the alloy!

[19] $\int_0^\infty dz \cdot \chi^4(z) = \frac{3b}{16}$ from Table 23.2.

[20] Here κ_b is an inverse penetration depth that depends on the conduction band offset ΔE_c. b is the only variational parameter. $m_b^\star$ and $m_w^\star$ are the effective masses in the barrier and well respectively. P_b is the fraction of mobile carriers that overlap with the alloy scattering sites in the barrier. See Exercise 23.4.

23.8 Interface scattering

Quantum confined electrons are scattered by the roughness of the barrier/well interface as indicated in Fig. 23.30. Such scattering is pictured in two complementary ways: via an electric field at the interface, or as fluctuations of the ground state energy eigenvalue of the conduction band in the 2D plane due to fluctuations in the well thickness. We discuss the eigenvalue method first.

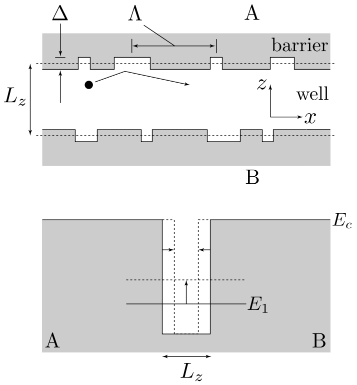


Fig. 23.30 Roughness in a quantum well thickness L_z causes local variations of the ground state energy E_1, which scatter electrons moving in the $x-y$ plane.

For a quantum well of average thickness L_z, let the actual thickness at a point (x,y) in the 2D plane be $L_z(x,y) = L_z + \Delta$, where $\Delta = \Delta(x,y)$ is the function that models the roughness. The ground state energy of the quantum well will also vary with (x,y), leading to the scattering potential and a corresponding matrix element obtained by using an infinite-well ground state and $\rho = \sqrt{x^2+y^2}$:

$$E_1(x,y) = \frac{h^2}{8m^\star L_z(x,y)^2} = \frac{h^2}{8m^\star L_z^2(1+\frac{\Delta}{L_z})^2} \approx \frac{h^2}{8m^\star L_z^2}(1-2\frac{\Delta}{L_z})$$

$$W(x,y) = -\frac{h^2}{4m^\star L_z^3}\Delta \implies W_{uns.}(q_\|) = -\frac{h^2}{4m^\star L_z^3}\int \frac{d^2r}{L^2} e^{i\mathbf{q}_\|\cdot\vec{\rho}}\Delta(x,y)$$

$$\implies |W_{uns.}(q_\|)|^2 = \frac{h^4}{16(m^\star)^2L_z^6}\int\int \frac{d^2r}{L^2}\frac{d^2r'}{L^2} e^{i\mathbf{q}_\|\cdot(\vec{\rho}-\vec{\rho}')}\Delta(\rho)\Delta(\rho')$$

$$= \frac{h^4}{16(m^\star)^2L_z^6L^4}\int_0^\infty d\rho\int_0^\infty d\rho'\cdot\rho\cdot\rho'\cdot\Delta(\rho)\cdot\Delta(\rho')\times \underbrace{\int_0^{2\pi} d\phi e^{iq_\|\rho\cos\phi}}_{2\pi J_0(q_\|\rho)}\underbrace{\int_0^{2\pi} d\phi' e^{iq_\|\rho'\cos\phi'}}_{2\pi J_0(q_\|\rho')}$$

$$= \frac{\pi^2h^4}{4(m^\star)^2L_z^6L^4}\int_{\rho=0}^\infty d\rho\int_{\rho'=0}^\infty d\rho'\cdot\rho\cdot\rho'\cdot[\Delta(\rho)\cdot\Delta(\rho')]\cdot J_0(q_\|\rho)\cdot J_0(q_\|\rho'). \tag{23.64}$$

The squared matrix element thus depends on the spatial correlation function $\Delta(\rho)\cdot\Delta(\rho')$. For an exponential correlation[21] the squared matrix element is exactly evaluated, leading to the scattering rate

$$\Delta(\rho)\Delta(\rho') = \delta^2 e^{-(\frac{\rho-\rho'}{\Lambda})} \implies |W_{uns.}(q_\|)|^2 = \frac{\pi^2h^4}{4(m^\star)^2L_z^6L^4}\frac{\delta^2\Lambda^4}{[1+(q_\|\Lambda)^2]^3}$$

$$\implies |W_{scr.}(q_\|)|^2 = \frac{\pi^2h^4}{4(m^\star)^2L_z^6L^4}\frac{\delta^2\Lambda^4}{[1+\frac{q_s}{2k_\|}P_1(q_\|)]^2[1+(q_\|\Lambda)^2]^3}$$

$$\implies \frac{1}{\tau_m(k_\|)} = \frac{N_{imp}L^2m^\star}{2\pi(\hbar k_\|)^3}\int_0^{2k_\|} dq_\| \frac{q_\|^2|W_{scr.}(q_\|)|^2}{\sqrt{1-(\frac{q_\|}{2k_\|})^2}}, \text{ where } N_{imp} = \frac{L^2}{\Lambda^2}$$

$$\frac{1}{\tau_m(k_\|)} = \frac{2\times 4\pi^5\hbar\delta^2\Lambda^2}{m^\star L_z^6}\int_0^1 du \frac{u^4}{\sqrt{1-u^2}[u+\frac{q_s}{2k_\|}P_1(u)]^2[1+(2k_\|\Lambda u)^2]^3}. \tag{23.65}$$

---

[21] The correlation functions $\Delta(x,y)$ are characterized by two length scales: thickness variations δ in the z-direction, and a correlation length Λ in the ρ-direction. This is illustrated in Fig. 23.30. With highly controlled crystal growth, δ can be limited to a monolayer of the crystal, and Λ can be 10s of lattice constants. See Exercise 23.6 for the evaluation of the matrix elements for the correlation functions.

A Gaussian correlation also leads to an exact matrix element, giving

$$\Delta(\rho)\Delta(\rho') = \delta^2 e^{-(\frac{\rho-\rho'}{\Lambda})^2} \implies |W_{uns.}(q_{\parallel})|^2 = \frac{\pi^6\hbar^4}{(m^\star)^2 L_z^6 L^4} \frac{\delta^2\Lambda^4 e^{-\frac{(q_{\parallel}\Lambda)^2}{2}}}{4}$$

$$\implies \frac{1}{\tau_m(k_{\parallel})} = \frac{4\pi^5\hbar\delta^2\Lambda^2}{m^\star L_z^6} \int_0^1 du \frac{u^4 e^{-2k_{\parallel}^2\Lambda^2u^2}}{\sqrt{1-u^2}[u+\frac{q_s}{2k_{\parallel}}P_1(u)]^2}. \tag{23.66}$$

Notice both scattering rates depend strongly on the width of the quantum well as $\sim 1/L_z^6$. This sixth-power law dependence causes severe degradation of mobility in very thin wells, falling as $\mu \sim L_z^6$.

Let us consider the second point of view, in which the scattering potential is caused because of the potential variation caused by the rough interface coupled with an electric field at the junction. Since the net potential experienced by electrons is $V(z,\Delta) = V(z) - F(z)\Delta$, the perturbation potential is $W(x,y) = qF(z)\Delta(x,y)$, where $F(z)$ is the electric field at z. Defining an effective "z-averaged" field $F_{eff} = \int dz\chi^2(z)F(z)$, for a Gaussian correlation function the matrix element and scattering rate are obtained similar to Equation 23.66:

$$F_{eff} = \int dz\chi^2(z)F(z) \implies |W_{scr.}(q_{\parallel})|^2 = \frac{\pi^2q^2F_{eff}^2\delta^2\Lambda^4}{L^4} \frac{e^{-\frac{q_{\parallel}^2\Lambda^2}{2}}}{[1+\frac{q_s}{q_{\parallel}}P_1(q_{\parallel})]^2}$$

$$\implies \frac{1}{\tau_m(k_{\parallel})} = \frac{4\pi m^\star q^2\delta^2\Lambda^2F_{eff}^2}{\hbar^3} \int_0^1 du \frac{u^4 e^{-2k_{\parallel}^2\Lambda^2u^2}}{\sqrt{1-u^2}[u+\frac{q_s}{2k_{\parallel}}P_1(u)]^2}$$

$$= \frac{4\pi m^\star q^4\delta^2\Lambda^2(N_dx_d+\eta n_{2deg})^2}{\epsilon_s^2\hbar^3} \int_0^1 du \frac{u^4 e^{-2k_{\parallel}^2\Lambda^2u^2}}{\sqrt{1-u^2}[u+\frac{q_s}{2k_{\parallel}}P_1(u)]^2}. \tag{23.67}$$

where the effective electric field $F_{eff} = \frac{q}{\epsilon_s}(N_dx_d + \eta n_{2deg})$ by Gauss's law, where N_dx_d is the integrated depletion charge of a MOSFET, with N_d the doping density and x_d the depletion depth. Here n_{2deg} is the sheet density of the mobile 2D carriers, with $\eta = \frac{1}{2}$ or $\eta = \frac{11}{32}$ used to model different 2DEG scenarios. In HEMTs where there are no depletion charges, $N_d = 0$ and $\eta = \frac{11}{32}$ is often used, and the fit to the δ and Λ is typically empirical. Since n_s can be varied by a gate voltage, bounds of δ and Λ can be determined from experiment if interface roughness scattering is the dominant scattering event.

Fig. 23.31 uses a Gaussian correlation function and illustrates several aspects of interface roughness scattering in semiconductor heterostructures. The population of carriers moves closer to the rough interface as the density increases, degrading the mobility. For degenerate 2DEGs replacing $k_{\parallel} \to k_F$ in Equations 23.67, 23.31, and 23.64, using $\mu_{2d}^{IR} \sim \frac{q\tau_m(k_F)}{m^\star}$ leads to the result that due to interface roughness scattering the 2DEG mobility drops roughly as $\mu_{2d}^{IR} \sim 1/n_{2deg}^2$.

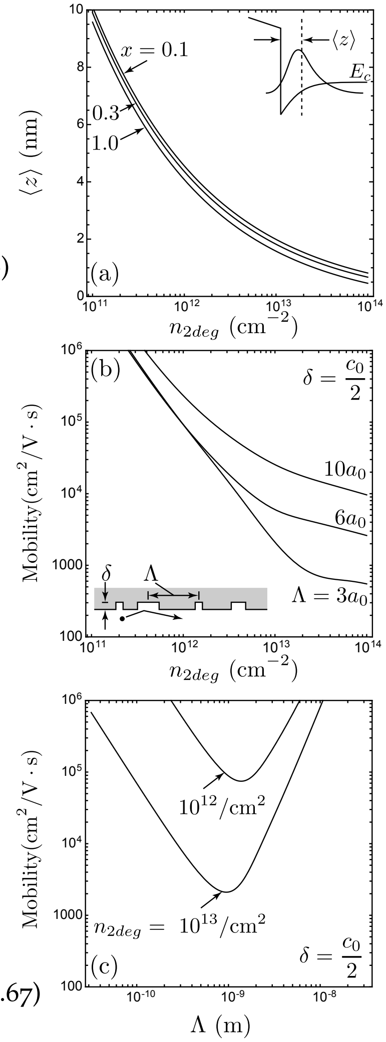


Fig. 23.31 Interface roughness scattering for 2D electrons at an Al$_x$Ga$_{1-x}$N/GaN heterojunction of c-axis lattice constant c_0 and c-plane lattice constant a_0. (a) The average distance of 2D carriers from a heterointerface as a function of the 2D electron gas concentration for various barrier alloy compositions. The electrons move closer to the interface at high densities. (b) Mobility due to interface roughness scattering lowers significantly with increase in carrier density. A smoother interface with higher Λ helps increase the mobility. (c) The mobility is the lowest when the electron Fermi wavelength λ_F matches the correlation length Λ.

23.9 Phonon scattering

The scattering potentials experienced by electrons from acoustic and optical phonons were discussed in Chapter 22. Before applying Fermi's golden rule to evaluate the individual electron scattering rates for 3D and 2D carriers, a physically transparent general result can be stated easily: all phonon scattering rates take the form

$$\frac{1}{\tau_m(\mathbf{k})} = \underbrace{C(\mathbf{k})}_{\text{el-ph. coupling coeff.}} \times \underbrace{\omega}_{\text{phonon frequency}} \times \underbrace{N_\omega}_{\text{phonon number}}, \qquad (23.68)$$

where momentum-dependent $C(\mathbf{k})$ is a dimensionless measure of the electron-phonon coupling, ω is the phonon frequency, and N_ω is the phonon number. For phonon absorption $N_\omega = N_0(\omega)$, and for phonon emission $N_\omega = N_0(\omega) + 1$, where $N_0(\omega)$ is the Bose–Einstein occupation function for the frequency ω of the respective phonon dispersion relation (acoustic or optical). This generalized form emphasizes that the frequency of scattering is related to the vibrational frequency of the crystal via ω, and the temperature via N_ω.[22] Written out explicitly for absorption and emission processes, Equation 23.68 is

$$\frac{1}{\tau_m(\mathbf{k})} = C_{abs}(\mathbf{k}) \cdot \omega \cdot (N_\omega) + C_{em}(\mathbf{k}) \cdot \omega \cdot (N_\omega + 1). \qquad (23.69)$$

The corresponding coefficients for absorption and emission are listed in Table 23.6 for 3D carriers. While the absorption and emission coefficients are nearly identical for quasi-elastic acoustic phonon scattering, they differ for inelastic optical phonon scattering. In the following, we will **(a)** first quantitatively evaluate the scattering rates for 3D carriers individually for each phonon branch, **(b)** re-do them for 2D carriers, and then **(c)** compare their relative importance with all results at hand.

The perturbation potential $W(r,t)$ experienced by electrons due to phonons discussed in Chapter 22 take the form $W_\pm(\mathbf{r},t) = K_\beta \cdot u_\beta \cdot e^{\pm i(\vec{\beta}\cdot\vec{r} - \omega_\beta t)} = W_0(\mathbf{r}) e^{\mp i\omega_\beta t}$, where the spatial part $W_0(\mathbf{r}) = K_\beta \cdot u_\beta \cdot e^{\pm i\vec{\beta}\cdot\vec{r}}$ is split off from the temporal part $e^{\mp i\omega_\beta t}$. The spatial part enters the matrix element, and the temporal part enters the energy conservation relation of Fermi's golden rule. The $+$ sign is for phonon absorption and $-$ is for phonon emission. The common factor for all scattering mechanisms is the vibration amplitude of the phonon wave u_β, whose square for phonon absorption process is $|u_\beta^{abs}|^2 = \frac{\hbar}{2\rho L^3 \omega_\beta}(N_\beta)$ and for phonon emission process is $|u_\beta^{em}|^2 = \frac{\hbar}{2\rho L^3 \omega_\beta}(N_\beta + 1)$. Emission is characterized by the sum of the stimulated ($\propto N_\beta$) and spontaneous ($\propto 1$) emission processes. These are listed in Table 23.5. This table also includes the ranges of phonon wavevectors for the absorption and emission processes, for the acoustic and optical modes, which result from the requirements of energy and momentum conservation in the electron-phonon interaction as was discussed in Chapter 22.

Table 23.5 Phonon amplitudes and ranges. The volume of the crystal is L^3, the mass density ρ, phonon dispersions are ω_β where β is the phonon wavevector and ω_β the corresponding frequency, N_β is the Bose–Einstein occupation function, v_s is the sound velocity, and $\hbar\omega_0$ is the optical phonon energy. $E_\mathbf{k}$ is the energy and $\hbar k$ the momentum of an electron in state k and effective mass $m^\star$.

	$\lvert$Amplitude$\rvert^2$	
$\lvert u_\beta^{abs}\rvert^2$	$\frac{\hbar(N_\beta)}{2\rho L^3\omega_\beta}$	
$\lvert u_\beta^{em}\rvert^2$	$\frac{\hbar(N_\beta+1)}{2\rho L^3\omega_\beta}$	
	Acoustic Absorption	Emission
$\frac{\beta_{max}}{k}$	0	$2(1 - \frac{m^\star v_s}{\hbar k})$
$\frac{\beta_{min}}{k}$	$2(1 + \frac{m^\star v_s}{\hbar k})$	0
	Optical Absorption	Emission
$\frac{\beta_{max}}{k}$	$\sqrt{1 + \frac{\hbar\omega_0}{E_\mathbf{k}}} + 1$	$1 + \sqrt{1 - \frac{\hbar\omega_0}{E_\mathbf{k}}}$
$\frac{\beta_{min}}{k}$	$\sqrt{1 + \frac{\hbar\omega_0}{E_\mathbf{k}}} - 1$	$1 - \sqrt{1 - \frac{\hbar\omega_0}{E_\mathbf{k}}}$

---

[22] Consider for example the technologically important compound semiconductors GaAs ($m_c^\star = 0.063 m_e, \hbar\omega_{LO} = 36$ meV) and GaN ($m_c^\star = 0.22 m_e, \hbar\omega_{LO} = 92$ meV). The highest measured room temperature electron mobilities in them are $\mu \sim 8000$ and $\mu \sim 2000$ $\frac{\text{cm}^2}{\text{V}\cdot\text{s}}$ respectively. Then the scattering rates $\frac{1}{\tau_m} \sim \frac{q}{\mu m_{eff}}$ are 3.5×10^{12}/s for GaAs and 4×10^{12} for GaN. The dominant scattering mechanism at room temperature in high purity samples of these semiconductors is due to the absorption of polar optical phonons. For GaAs a LO phonon frequency $\omega_{LO} = 5.5 \times 10^{13}$/s and Bose occupation function $N_0(\omega_{LO}) = 0.33$ leads to a coupling coefficient $C(\mathbf{k}) \sim 0.19$. For GaN, the frequency $\omega_{LO} = 1.4 \times 10^{14}$/s and $N_0(\omega_{LO}) = 0.03$ yields a nearly 5× stronger electron-phonon coupling coefficient at $C(\mathbf{k}) \sim 0.97$. Note that here $C(\mathbf{k})$ is of order unity.

ble 23.6 Phonon scattering matrix elements and momentum relaxation rates. The total scattering rate $\mathbf{k})$ $= C_{abs}(\mathbf{k}) \cdot \omega \cdot (N_\omega) + C_{em}(\mathbf{k}) \cdot \omega \cdot (N_\omega + 1)$ is the sum of absorption and emission. The rates characterized by the dimensionless absorption and emission coefficients $C_{abs}(\mathbf{k})$ and $C_{em}(\mathbf{k})$ that pend on the electron momentum. The dimensionless Frohlich coupling parameter for polar optical onon scattering is $\alpha_{LO} = \frac{\frac{1}{2}\frac{q^2 k_{LO}}{4\pi}(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})}{\hbar\omega_0}$.

Type	$W_0(\beta) = K_\beta u_\beta$	$\|K_\beta\|^2$	$C_{abs}(\mathbf{k})$	$C_{em}(\mathbf{k})$
ADP	$(iD_a\beta)u_\beta$	$\beta^2 D_a^2$	$\frac{m^\star D_a^2}{\pi\hbar^2\rho v_s^2}k$	$\approx C_{abs}(\mathbf{k})$
PZ	$(\frac{qe_{pz}}{\epsilon_{dc}})u_\beta$	$(\frac{qe_{pz}}{\epsilon_{dc}})^2$	$\frac{m^\star q^2 e_{pz}^2}{2\pi\hbar^2\rho v_s^2\epsilon_{dc}^2 k}$	$\approx C_{abs}(\mathbf{k})$
ODP	$(D_{op})u_\beta$	D_{op}^2	$\frac{m^\star}{2\pi\rho}(\frac{D_{op}}{\hbar\omega_0})^2 \times$ $\sqrt{\frac{2m^\star(E_{\mathbf{k}}+\hbar\omega_0)}{\hbar^2}}$	$\frac{m^\star}{2\pi\rho}(\frac{D_{op}}{\hbar\omega_0})^2 \times$ $\sqrt{\frac{2m^\star(E_{\mathbf{k}}-\hbar\omega_0)}{\hbar^2}}$
POP	$\frac{q\omega_0}{i\beta}\sqrt{\rho(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})}u_\beta$	$\frac{\rho q^2\omega_0^2}{\beta^2}(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})$	$\alpha_{LO} \times \frac{k_{LO}}{k} \times$ $\Big(\sqrt{1 + \frac{\hbar\omega_0}{E_k}}$ $- \frac{\hbar\omega_0}{E_k}\sinh^{-1}\sqrt{\frac{E_k}{\hbar\omega_0}}\Big)$	$\alpha_{LO} \times \frac{k_{LO}}{k} \times$ $\Big(\sqrt{1 - \frac{\hbar\omega_0}{E_k}}$ $+ \frac{\hbar\omega_0}{E_k}\sinh^{-1}\sqrt{\frac{E_k}{\hbar\omega_0} - 1}\Big)$

The squared matrix element for scattering of electrons in 3D is then

$$\langle f|W_0(\mathbf{r})|i\rangle = \int d^3r \cdot \psi_f^\star(\mathbf{r}) \cdot W_0(\mathbf{r}) \cdot \psi_i(\mathbf{r})$$

$$= \int d^3r[\frac{e^{-i\mathbf{k}'\cdot\mathbf{r}}}{\sqrt{L^3}}u_{\mathbf{k}'}^\star(\mathbf{r})] \cdot [K_\beta \cdot u_\beta \cdot e^{\pm i\vec{\beta}\cdot\vec{r}}] \cdot [\frac{e^{+i\mathbf{k}\cdot\mathbf{r}}}{\sqrt{L^3}}u_{\mathbf{k}}(\mathbf{r})]$$

$$= K_\beta u_\beta I_{\mathbf{k},\mathbf{k}'}\underbrace{\int\frac{d^3r}{L^3}e^{i(\mathbf{k}-\mathbf{k}'\pm\vec{\beta})\cdot\mathbf{r}}}_{\delta_{\mathbf{k}',\mathbf{k}\pm\vec{\beta}}} \implies \langle f|W_0(\mathbf{r})|i\rangle \approx K_\beta u_\beta \delta_{\mathbf{k}',\mathbf{k}\pm\vec{\beta}}$$

$$\implies \boxed{|\langle f|W_0(\mathbf{r})|i\rangle|^2 = |K_\beta|^2 \cdot |u_\beta|^2 \cdot \delta_{\mathbf{k}',\mathbf{k}\pm\vec{\beta}},} \tag{23.70}$$

where[23] the individual scattering potentials of the four phonon modes and the corresponding coupling terms $|K_\beta|^2$ are listed in Table 23.6. Fermi's golden rule then gives the scattering rates

$$S(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar} \cdot |\langle f|W_0(\mathbf{r})|i\rangle|^2 \cdot \delta[E_f - (E_i \pm \hbar\omega_\beta)] \implies$$

$$S_{abs}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}|K_\beta|^2 \cdot |u_\beta^{abs}|^2 \cdot \delta_{\mathbf{k}',\mathbf{k}+\vec{\beta}} \cdot \delta[E_f - (E_i + \hbar\omega_\beta)], \text{ and}$$

$$S_{em}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}|K_\beta|^2 \cdot |u_\beta^{em}|^2 \cdot \delta_{\mathbf{k}',\mathbf{k}-\vec{\beta}} \cdot \delta[E_f - (E_i - \hbar\omega_\beta)]$$

$$\text{leading to } \frac{1}{\tau_m(\mathbf{k})} = \sum_\beta\sum_{\mathbf{k}'}(1 - \frac{\mathbf{k}'\cdot\mathbf{k}}{|\mathbf{k}|^2}) \cdot S(\mathbf{k},\mathbf{k}') \tag{23.71}$$

[23] Note that the dimensionless Kronecker delta function is equal to its square. The squared matrix element of Equation 23.70 has units of (energy)^2, $|u_\beta|^2$ is of (length)^2 units, and K_β is of energy/length units. The Kronecker-delta enforces $\hbar\mathbf{k}' = \hbar\mathbf{k} \pm \hbar\vec{\beta}$ which is momentum conservation for normal scattering. If Umklapp scattering is not negligible, the momentum conservation relation is $\hbar\mathbf{k}' = \hbar\mathbf{k} \pm \hbar\vec{\beta} \pm \hbar\mathbf{G}$, where $\mathbf{G}$ is a reciprocal lattice vector. Umklapp processes are typically negligible for electrical conductivity of semiconductors.

separately evaluated for absorption and emission processes. This recipe is now applied to calculate the individual momentum scattering rates[24].

Longitudinal acoustic phonons, 3D carriers: From Table 23.6, for acoustic deformation potential (ADP) scattering, $K_\beta = i\beta D_a$. The absorption rate is evaluated first. The geometry of Fig. 23.32 is used, and the steps are written out in explicit detail:

$$\frac{1}{\tau_{m,abs}^{ADP}(\mathbf{k})} = \sum_{\beta}\sum_{\mathbf{k}'}(1 - \frac{\mathbf{k}'\cdot\mathbf{k}}{|\mathbf{k}|^2})\cdot S_{abs}(\mathbf{k},\mathbf{k}')$$

$$= \sum_{\beta}\sum_{\mathbf{k}'}(1 - \frac{\mathbf{k}'\cdot\mathbf{k}}{|\mathbf{k}|^2})\frac{2\pi}{\hbar}|K_\beta|^2\cdot|u_\beta^{abs}|^2\cdot\delta_{\mathbf{k}',\mathbf{k}+\vec{\beta}}\cdot\delta[E_f - (E_i + \hbar\omega_\beta)]$$

$$= \frac{2\pi}{\hbar}\sum_{\beta}(1 - \frac{(\mathbf{k}+\vec{\beta})\cdot\mathbf{k}}{|\mathbf{k}|^2})\cdot\beta^2 D_a^2\cdot\frac{\hbar N_\beta}{2\rho L^3\omega_\beta}\cdot\delta[E_f - (E_i + \hbar\omega_\beta)]$$

$$= \frac{\pi D_a^2}{\rho L^3}\sum_{\beta}(-\frac{\beta}{k}\cos\alpha)\cdot\beta^2\cdot\frac{k_b T}{\hbar\omega_\beta^2}\cdot\delta[\frac{\hbar^2|\mathbf{k}+\vec{\beta}|^2}{2m^\star} - (\frac{\hbar^2|\mathbf{k}|^2}{2m^\star} + \hbar v_s\beta)]$$

$$= \frac{\pi D_a^2 k_b T}{\hbar\rho v_s^2 L^3}\sum_{\beta}(-\frac{\beta}{k}\cos\alpha)\cdot\frac{\beta^2}{\beta^2}\cdot\frac{2m^\star}{\hbar^2}\delta[\beta^2 + 2\beta k\cos\alpha - 2\beta\underbrace{\frac{m^\star v_s}{\hbar}}_{k_s}]$$

$$= \frac{2\pi m^\star D_a^2 k_b T}{\hbar^3\rho v_s^2 L^3 k}\sum_{\beta}(-\beta\cos\alpha)\cdot\delta[\beta^2 + 2\beta k\cos\alpha - 2\beta k_s], \tag{23.72}$$

where the sum is now over all allowed phonon wavevectors $\vec{\beta}$. For 3D carriers, this converts to an integral in spherical coordinates:

$$= \frac{2\pi m^\star D_a^2 k_b T}{\hbar^3\rho v_s^2 L^3 k}\int\frac{d^3\beta}{\frac{(2\pi)^3}{L^3}}(-\beta\cos\alpha)\cdot\delta[\beta^2 + 2\beta k\cos\alpha - 2\beta k_s]$$

$$= \frac{2\pi m^\star D_a^2 k_b T}{\hbar^3\rho v_s^2(2\pi)^3 k}\int\beta^2\sin\alpha\, d\beta\, d\alpha\, d\phi(-\beta\cos\alpha)\cdot\delta[\beta^2 + 2\beta k\cos\alpha - 2\beta k_s]$$

$$= \frac{2\pi m^\star D_a^2 k_b T}{\hbar^3\rho v_s^2(2\pi)^3 k}\underbrace{\int_0^{2\pi}d\phi}_{2\pi}\int_0^{2k(1-\frac{k_s}{k})}d\beta\cdot\beta^3\times$$

$$\cdot\int_{\alpha=\frac{\pi}{2}}^{\pi}d(\cos\alpha)\cos\alpha\cdot\delta[\beta^2 + 2\beta k\cos\alpha - 2\beta k_s]$$

$$= \frac{m^\star D_a^2 k_b T}{2\pi\hbar^3\rho v_s^2 k}\int_0^{2k(1-\frac{k_s}{k})}d\beta\cdot\beta^3\cdot\underbrace{\int_0^{-1}du\cdot u\cdot\delta[\beta^2 + 2\beta k u - 2\beta k_s]}_{=\frac{1}{4k^2}\text{ for }\frac{m^\star v_s}{\hbar k}<<1}$$

$$= \frac{m^\star D_a^2 k_b T}{2\pi\hbar^3\rho v_s^2 k}\cdot\underbrace{\frac{1}{4k^2}\int_0^{2k(1-\frac{k_s}{k})}d\beta\cdot\beta^3}_{=k^2\text{ for }\frac{m^\star v_s}{\hbar k}<<1} \implies \frac{1}{\tau_{m,abs}^{ADP}(\mathbf{k})} = \frac{m^\star D_a^2 k_b T}{2\pi\hbar^3\rho v_s^2}\cdot k, \tag{23.73}$$

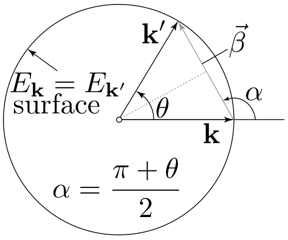

$\mathbf{k}'$
$\vec{\beta}$
$E_{\mathbf{k}} = E_{\mathbf{k}'}$
surface
θ
α
$\mathbf{k}$
$\alpha = \frac{\pi+\theta}{2}$

Fig. 23.32 Geometry of electron scattering from acoustic phonons, assumed quasi-elastic. The initial and final electron wavevectors $(\mathbf{k}, \mathbf{k}')$ lie on a constant energy sphere for 3D transport. Note that the angle between the phonon wavevector $\vec{\beta}$ and the initial electron wavevector $\mathbf{k}$ is α, which is related to the scattering angle θ by $\alpha = (\pi+\theta)/2$. So when $\theta : 0 \to \pi$, $\alpha : \frac{\pi}{2} \to \pi$.

[24] The corresponding quantum relaxation rate $\frac{1}{\tau_q(\mathbf{k})} = \sum_\beta\sum_{\mathbf{k}'}S(\mathbf{k},\mathbf{k}')$ and energy relaxation rate $\frac{1}{\tau_q(\mathbf{k})} = \sum_\beta\sum_{\mathbf{k}'}(1 - \frac{E'_{\mathbf{k}}}{E_{\mathbf{k}}})S(\mathbf{k},\mathbf{k}')$ are evaluated similarly by using the appropriate weighting factors of $S(\mathbf{k},\mathbf{k}')$ in the sums.

where the appropriate angular and radial limits set by energy and momentum conservation requirements are enforced to obtain the analytical result. Because for low-energy acoustic phonons $\frac{k_b T}{\hbar v_s \beta} >> 1$ even at very low temperatures, $N_\beta \approx \frac{k_b T}{\hbar v_s \beta}$ is very nearly equal to $N_\beta + 1 \approx \frac{k_b T}{\hbar v_s \beta} + 1$. The total scattering rate therefore is roughly twice of the absorption rate, leading to a mobility[25]

$$\frac{1}{\tau^{ADP}_{m,em}(\mathbf{k})} \approx \frac{1}{\tau^{ADP}_{m,bas}(\mathbf{k})} \implies \boxed{\frac{1}{\tau^{ADP}_{m}(\mathbf{k})} \approx \frac{2}{\tau^{ADP}_{m,abs}(\mathbf{k})} = \frac{m^\star D_a^2 k_b T}{\pi \hbar^3 \rho v_s^2} \cdot k}$$
$$\implies \boxed{\mu^{ADP}_{3d} = \frac{\sqrt{\frac{8\pi}{9}} \cdot q\hbar^4 \rho v_s^2}{(m^\star)^{\frac{5}{2}} D_a^2 (k_b T)^{\frac{3}{2}}} \cdot \frac{\ln(1+e^\eta)}{F_{\frac{1}{2}}(\eta)}} \approx \frac{\sqrt{\frac{8\pi}{9}} \cdot q\hbar^4 \rho v_s^2}{(m^\star)^{\frac{5}{2}} D_a^2 (k_b T)^{\frac{3}{2}}}. \tag{23.74}$$

Piezoelectric acoustic phonons, 3D carriers: From Table 23.6, for acoustic piezoelectric phonon (PZ) scattering, $K_\beta = \frac{q e_{pz}}{\epsilon_{dc}}$, with which we can proceed in a way similar to the calculation of ADP rates to compute the scattering rate[26]. The momentum scattering rate and mobility limited by piezoelectric phonon scattering is thus given by

$$\frac{1}{\tau^{PZ}_{m,abs}(\mathbf{k})} = \sum_\beta \sum_{\mathbf{k}'} (1 - \frac{\mathbf{k}' \cdot \mathbf{k}}{|\mathbf{k}|^2}) \cdot S_{abs}(\mathbf{k}, \mathbf{k}')$$
$$= \frac{2\pi}{\hbar} \sum_\beta (1 - \frac{(\mathbf{k} + \vec{\beta}) \cdot \mathbf{k}}{|\mathbf{k}|^2}) \cdot \frac{q^2 e_{pz}^2}{\epsilon_{dc}^2} \cdot \frac{\hbar N_\beta}{2\rho L^3 \omega_\beta} \cdot \delta[E_f - (E_i + \hbar\omega_\beta)]$$
$$= \frac{2\pi m^\star q^2 e_{pz}^2 k_b T}{\hbar^3 \rho v_s^2 \epsilon_{dc}^2 L^3} \sum_\beta (-\frac{\beta}{k} \cos\alpha) \cdot \frac{1}{\beta^2} \cdot \delta[\beta^2 + 2\beta k \cos\alpha - 2\beta k_s]$$
$$= \frac{m^\star q^2 e_{pz}^2 k_b T}{2\pi \hbar^3 \rho v_s^2 \epsilon_{dc}^2 k} \int \beta^2 \sin\alpha d\beta d\alpha (-\frac{\cos\alpha}{\beta}) \cdot \delta[\beta^2 + 2\beta k \cos\alpha - 2\beta k_s]$$
$$= \frac{m^\star q^2 e_{pz}^2 k_b T}{2\pi \hbar^3 \rho v_s^2 \epsilon_{dc}^2 k} \int_0^{2k} d\beta \cdot \beta \int_{\alpha=\frac{\pi}{2}}^{\pi} d(\cos\alpha) \cos\alpha \cdot \delta[\beta^2 + 2\beta k \cos\alpha - 2\beta k_s]$$
$$= \frac{m^\star q^2 e_{pz}^2 k_b T}{2\pi \hbar^3 \rho v_s^2 \epsilon_{dc}^2 k} \cdot \frac{1}{4k^2} \int_0^{2k} d\beta \cdot \beta \implies \frac{1}{\tau^{PZ}_{m,abs}(\mathbf{k})} = \frac{m^\star q^2 e_{pz}^2 k_b T}{4\pi \hbar^3 \rho v_s^2 \epsilon_{dc}^2} \cdot \frac{1}{k}$$
$$\implies \boxed{\frac{1}{\tau^{PZ}_{m}(\mathbf{k})} \approx \frac{2}{\tau^{PZ}_{m,abs}(\mathbf{k})} = \frac{m^\star q^2 e_{pz}^2 k_b T}{2\pi \hbar^3 \rho v_s^2 \epsilon_{dc}^2} \cdot \frac{1}{k}}$$
$$\implies \boxed{\mu^{PZ}_{3d} = \frac{\frac{16\sqrt{2\pi}}{3} \cdot \rho v_s^2 \hbar^2 \epsilon_{dc}^2}{q (m^\star)^{\frac{3}{2}} e_{pz}^2 (k_b T)^{\frac{1}{2}}} \cdot \frac{F_1(\eta)}{F_{\frac{1}{2}}(\eta)}} \approx \frac{\frac{16\sqrt{2\pi}}{3} \cdot \rho v_s^2 \hbar^2 \epsilon_{dc}^2}{q (m^\star)^{\frac{3}{2}} e_{pz}^2 (k_b T)^{\frac{1}{2}}}, \tag{23.75}$$

which decreases as $\mu^{PZ}_{3d} \sim T^{-\frac{1}{2}}$, slower than the ADP scattering.

---

[25] The ADP limited electron mobility in 3D therefore decreases with increasing temperature as $\mu^{ADP}_{3d} \sim T^{-\frac{3}{2}}$, a well-known result for elemental semiconductors such as silicon in which such phonon scattering dominates near room temperature in pure samples.

[26] Before doing so, we should pause and ask whether the same phonon mode ω_β indeed leads to the two independent forms of scattering with no interference? This is understood by noting that the scattering potentials are out of phase: $K^{ADP}_\beta = i\beta D_a$ is imaginary whereas $K^{PZ}_\beta = \frac{q e_{pz}}{\epsilon_{dc}}$ is real. Thus, the total matrix element squared $|K^{ADP}_\beta + K^{PZ}_\beta|^2 = |K^{ADP}_\beta|^2 + |K^{PZ}_\beta|^2$ is indeed the sum of the squares of the individual matrix elements, the two scattering mechanisms operate in parallel, and the rates add.

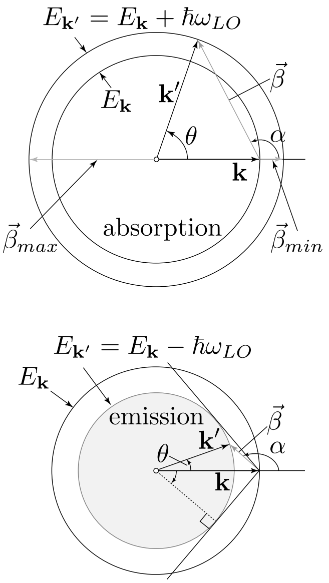


Fig. 23.33 Geometry of electron scattering from optical phonons. The characteristic wavevector k_{LO} is defined via the relation $(\hbar^2 k_{LO}^2)/(2m^\star) = \hbar\omega_{LO}$. **For phonon absorption,** the electron energy difference leads to $(k')^2 = k^2 + k_{LO}^2$. The angle α can take all values between $(0,\pi)$, but the phonon wavevectors range in magnitude from $\beta^{abs}_{min} = k' - k = k[\sqrt{1+(k_{LO}^2/k^2)} - 1]$ to $\beta^{abs}_{max} = k' + k = k[\sqrt{1+(k_{LO}^2/k^2)} + 1]$, as can be read off from the figure. **For phonon emission,** $(k')^2 = k^2 - k_{LO}^2$, which leads to $\beta^{em}_{min} = k - k' = k[1 - \sqrt{1-(k_{LO}^2/k^2)}]$ and $\beta^{em}_{max} = k + k' = k[1 + \sqrt{1-(k_{LO}^2/k^2)}]$. Clearly $E_{\mathbf{k}} \geq \hbar\omega_{LO}$, or equivalently $k \geq k_{LO}$ for an electron to be able to emit an optical phonon. The tangent from the tip of the vector $\mathbf{k}$ to the smaller circle defines a maximum angle θ, which implies that α ranges from π to a maximum value given by $\cos\alpha_{max} = k_{LO}/k$.

Optical deformation potential, 3D carriers: $K_\beta = D_{op}$ in units of eV/cm from Table 23.6 for optical deformation potential (ODP) scattering. Because the optical branch has $\omega = \omega_{LO}$, the scattering is inelastic, and we must accurately account for the limits on the magnitudes and angles of $\vec{\beta}$ placed by momentum and energy conservation relations. Fig. 23.33 shows the geometry for absorption and emission of optical phonons separately, defining the limits of the angles and phonon wavevectors that are allowed for each process. The scattering rates are

$$\frac{1}{\tau^{ODP}_{m,abs}(\mathbf{k})} = \sum_{\beta}\sum_{\mathbf{k}'}(1 - \frac{\mathbf{k}'\cdot\mathbf{k}}{|\mathbf{k}|^2}) \cdot S_{abs}(\mathbf{k},\mathbf{k}')$$

$$= \frac{2\pi}{\hbar}\sum_{\beta}(1 - \frac{(\mathbf{k}+\vec{\beta})\cdot\mathbf{k}}{|\mathbf{k}|^2}) \cdot D_{op}^2 \cdot \frac{\hbar N_{LO}}{2\rho L^3 \omega_{LO}} \cdot \delta[E_f - (E_i + \hbar\omega_{LO})]$$

$$= \frac{2\pi m^\star D_{op}^2 N_{LO}}{\hbar^2 \rho \omega_{LO} L^3}\sum_{\beta}(-\frac{\beta}{k}\cos\alpha) \cdot \delta[\beta^2 + 2\beta k\cos\alpha - k_{LO}^2]$$

$$= \frac{m^\star D_{op}^2 N_{LO}}{2\pi\hbar^2\rho\omega_{LO}k}\int \beta^2 \sin\alpha d\beta d\alpha \cdot (-\beta\cos\alpha) \cdot \delta[\beta^2 + 2\beta k\cos\alpha - k_{LO}^2]$$

$$= \frac{m^\star D_{op}^2 N_{LO}}{2\pi\hbar^2\rho\omega_{LO}k}\int_{\beta^{abs}_{min}}^{\beta^{abs}_{max}} d\beta \cdot \beta^3 \underbrace{\int_{\alpha=0}^{\pi} d(\cos\alpha)\cos\alpha \cdot \delta[\beta^2 + 2\beta k\cos\alpha - k_{LO}^2]}_{\frac{\beta^2 - k_{LO}^2}{4k^2\beta^2}}$$

$$= \frac{m^\star D_{op}^2 N_{LO}}{2\pi\hbar^2\rho\omega_{LO}k} \cdot k^2\sqrt{1 + \frac{k_{LO}^2}{k^2}}$$

$$\implies \frac{1}{\tau^{ODP}_{m,abs}(\mathbf{k})} = \frac{\pi D_{op}^2 N_{LO}}{2\rho\omega_{LO}} \cdot \underbrace{\frac{1}{2\pi^2}(\frac{2m^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{E_{\mathbf{k}} + \hbar\omega_{LO}}}_{g^{abs}_{3d}},$$

and similarly evaluating the emission rate, we get (see Exercise 23.7)

$$\frac{1}{\tau^{ODP}_{m,em}(\mathbf{k})} = \frac{\pi D_{op}^2 (N_{LO}+1)}{2\rho\omega_{LO}} \cdot \underbrace{\frac{1}{2\pi^2}(\frac{2m^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{E_{\mathbf{k}} - \hbar\omega_{LO}}}_{g^{em}_{3d}} \cdot \Theta(E_{\mathbf{k}} - \hbar\omega_{LO})$$

$$\implies \boxed{\frac{1}{\tau^{ODP}_{m}} = \frac{\pi D_{op}^2}{2\rho\omega_{LO}}[N_{LO}\cdot g^{abs}_{3d} + (N_{LO}+1)\cdot g^{em}_{3d}\cdot\Theta(E_{\mathbf{k}} - \hbar\omega_{LO})].} \tag{23.76}$$

Since the rate is not a simple polynomial function of energy, the mobility must be evaluated numerically. For non-degenerate carriers, the following relation is obtained[27] with $z = \hbar\omega_{LO}/k_b T$:

$$\mu^{ODP}_{3d} \approx \frac{\frac{4\sqrt{2\pi}}{3}e\hbar^2\rho(\hbar\omega_{LO})^{\frac{1}{2}}}{(m^\star)^{\frac{5}{2}}D_{op}^2 N_{LO}}(\frac{\hbar\omega_{LO}}{k_b T})^{\frac{5}{2}}\int_{y=0}^{\infty}\frac{dy \cdot y^{\frac{3}{2}} \cdot e^{-zy}}{\sqrt{y-1} + e^z\sqrt{y-1}\cdot\Theta(y-1)}. \tag{23.77}$$

[27] This is obtained directly by Equation 23.76 and the generalized mobility expression of Equation 23.34: try it! Note that for the Bose–Einstein occupation function, $(N_{LO}+1)/N_{LO} = e^z$.

Polar optical phonons, 3D carriers: From Table 23.6 we read off $K_\beta = \frac{C_0}{i\beta}$ where $C_0 = q\omega_0\sqrt{\rho(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})}$ is a material constant for polar optical phonon (POP) scattering. We note again the POP potential is orthogonal in the complex plane to the ODP potential, implying the two rates originating from the same phonon modes do not interfere, and operate in parallel: their rates add. The momentum relaxation rate[28] due to POP absorption is (see Fig. 23.34):

$$\frac{1}{\tau^{POP}_{m,abs}(\mathbf{k})} = \sum_{\beta}\sum_{\mathbf{k}'}(1 - \frac{\mathbf{k}'\cdot\mathbf{k}}{|\mathbf{k}|^2}) \cdot S_{abs}(\mathbf{k},\mathbf{k}')$$

$$= \frac{2\pi}{\hbar}\sum_{\beta}(1 - \frac{(\mathbf{k}+\vec{\beta})\cdot\mathbf{k}}{|\mathbf{k}|^2}) \cdot |\frac{C_0}{i\beta}|^2 \cdot \frac{\hbar N_{LO}}{2\rho L^3\omega_{LO}} \cdot \delta[E_f - (E_i + \hbar\omega_{LO})]$$

$$= \frac{2\pi m^\star C_0^2 N_{LO}}{\hbar^2\rho L^3\omega_{LO}}\sum_{\beta}(-\frac{\beta}{k}\cos\alpha)\cdot\frac{1}{\beta^2}\cdot\delta[\beta^2 + 2\beta k\cos\alpha - k_{LO}^2]$$

$$= \frac{m^\star C_0^2 N_{LO}}{2\pi\hbar^2\rho\omega_{LO}k}\int \beta^2\sin\alpha d\beta d\alpha \cdot (-\frac{\cos\alpha}{\beta})\cdot\delta[\beta^2 + 2\beta k\cos\alpha - k_{LO}^2]$$

$$= \frac{m^\star C_0^2 N_{LO}}{2\pi\hbar^2\rho\omega_{LO}k}\int_{\beta_{min}^{abs}}^{\beta_{max}^{abs}} d\beta\cdot\beta\underbrace{\int_{\alpha=0}^{\pi} d(\cos\alpha)\cos\alpha\cdot\delta[\beta^2 + 2\beta k\cos\alpha - k_{LO}^2]}_{\frac{\beta^2 - k_{LO}^2}{4k^2\beta^2}}$$

$$\frac{1}{\tau^{POP}_{m,abs}(\mathbf{k})} = \omega_{LO}\alpha_{LO}N_{LO}\frac{k_{LO}}{k}\Big[\sqrt{1 + (\frac{k_{LO}}{k})^2} - (\frac{k_{LO}}{k})^2\sinh^{-1}(\frac{k}{k_{LO}})\Big]$$

$$\frac{1}{\tau^{POP}_{m,abs}(E_{\mathbf{k}})} = \omega_{LO}\alpha_{LO}N_{LO}\sqrt{\frac{\hbar\omega_0}{E_{\mathbf{k}}}}\Big[\sqrt{1 + \frac{\hbar\omega_0}{E_{\mathbf{k}}}} - \frac{\hbar\omega_0}{E_{\mathbf{k}}}\sinh^{-1}\sqrt{\frac{E_{\mathbf{k}}}{\hbar\omega_0}}\Big]. \tag{23.78}$$

For POP emission, $0 \leq \alpha \leq \alpha_{max}$ with $\cos\alpha_{max} = \frac{k_{LO}}{k}$ gives the rates

$$\frac{1}{\tau^{POP}_{m,em}(\mathbf{k})} = \sum_{\beta}\sum_{\mathbf{k}'}(1 + \frac{\mathbf{k}'\cdot\mathbf{k}}{|\mathbf{k}|^2}) \cdot S_{em}(\mathbf{k},\mathbf{k}')$$

$$= \frac{m^\star C_0^2(N_{LO}+1)}{2\pi\hbar^2\rho\omega_{LO}k}\int_{\beta_{min}^{abs}}^{\beta_{max}^{abs}} d\beta\beta\int_{\alpha=0}^{\alpha_{max}} d(\cos\alpha)\cos\alpha\delta[\beta^2 - 2\beta k\cos\alpha + k_{LO}^2]$$

$$\frac{1}{\tau^{POP}_{m,em}(\mathbf{k})} = \omega_{LO}\alpha_{LO}(N_{LO}+1)\times$$

$$\frac{k_{LO}}{k}\cdot\Big[\sqrt{1 - (\frac{k_{LO}}{k})^2} + (\frac{k_{LO}}{k})^2\sinh^{-1}\sqrt{(\frac{k}{k_{LO}})^2 - 1}\Big]$$

$$\frac{1}{\tau^{POP}_{m,em}(E_{\mathbf{k}})} = \omega_{LO}\alpha_{LO}(N_{LO}+1)\times$$

$$\sqrt{\frac{\hbar\omega_0}{E_{\mathbf{k}}}}\cdot\Big[\sqrt{1 - \frac{\hbar\omega_0}{E_{\mathbf{k}}}} + \frac{\hbar\omega_0}{E_{\mathbf{k}}}\sinh^{-1}\sqrt{\frac{E_{\mathbf{k}}}{\hbar\omega_0} - 1}\Big]. \tag{23.79}$$

Fig. 23.34 These electron-polar optical phonon scattering rates were first obtained by Herbert Callen, a pioneer in statistical thermodynamics. Callen proved the fluctuation-dissipation theorem with T. Welton in 1951. It is one of the subject's most remarkable results that encompasses drift-diffusion relations, Brownian motion and Nyquist noise among other things.

---

[28] We emphasize that a momentum relaxation rate cannot be strictly defined for inelastic POP scattering because the BTE cannot be linearized for the general POP case. More on this point is discussed later. In the following derivation, we use the identity $\ln(x + \sqrt{x^2+1}) = \sinh^{-1}(x)$ and the dimensionless Frohlich coupling coefficient $\alpha_{LO} = \frac{1}{2}[\frac{q^2 k_{LO}}{4\pi}(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_{dc}})]/(\hbar\omega_0)$ to express the LO phonon scattering rate in a physically transparent way.

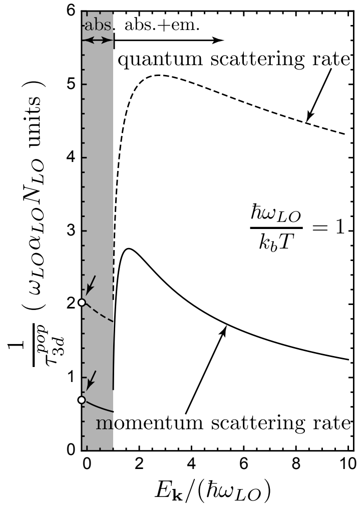


Fig. 23.35 Relative POP scattering rate strengths as a function of carrier energy.

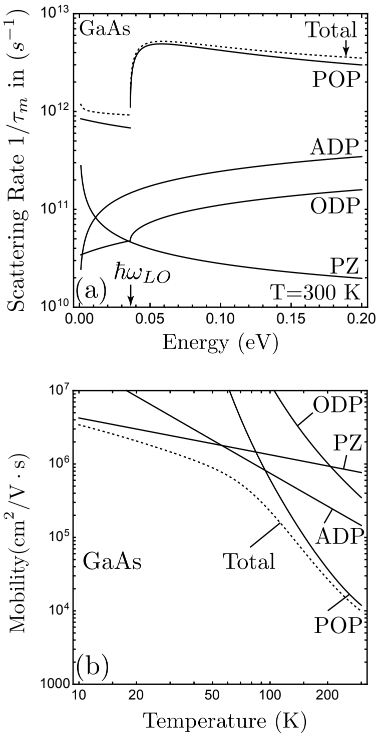


Fig. 23.36 (a) Scattering rates of phonons for GaAs at 300 K, and (b) phonon-limited electron mobility in 3D GaAs for $n = 10^{14}/\text{cm}^3$. Note the predominance of the POP scattering mechanism at room temperature.

Writing $x = (\hbar\omega_{LO})/E_{\mathbf{k}}$ and $z = (\hbar\omega_{LO})/(k_bT)$, the total polar optical phonon momentum relaxation rate is expressed compactly as

$$\frac{1}{\tau_m^{pop}} = \frac{1}{\tau_{m,abs}^{pop}} + \frac{1}{\tau_{m,em}^{pop}} = \omega_{LO}\alpha_{LO}N_{LO} \times \sqrt{x}\Big[\big(\sqrt{1+x} - x\sinh^{-1}\frac{1}{\sqrt{x}}\big) + e^z\big(\sqrt{1-x} + x\sinh^{-1}\sqrt{\frac{1}{x}-1}\big)\Theta(1-x)\Big], \tag{23.80}$$

where the Heaviside unit step $\Theta(1-x)$ ensures that for POP emission to occur $E_{\mathbf{k}} \geq \hbar\omega_{LO}$ (implying $x < 1$) because an electron cannot emit an optical phonon if it does not have kinetic energy at least equal to one optical phonon energy quantum. The same procedure gives the POP quantum relaxation rate:

$$\frac{1}{\tau_q^{pop}} = \omega_{LO}\alpha_{LO}N_{LO} \times 2\sqrt{x}[\sinh^{-1}\frac{1}{\sqrt{x}} + e^z\Theta(1-x)\sinh^{-1}\sqrt{\frac{1}{x}-1}]. \tag{23.81}$$

The momentum scattering rate is shown as the solid line in Fig. 23.35, and the quantum scattering rate as the dashed line. Note that momentum scattering rate by POP has a low-energy ($E_{\mathbf{k}} \to 0 \implies x \to \infty$) limit of $1/\tau_m^{POP} = \frac{2}{3} \times \omega_{LO}\alpha_{LO}N_{LO}$, and that the rate reaches a maximum slightly above $E_{\mathbf{k}} > \hbar\omega_{LO}$. The corresponding low-energy POP absorption quantum scattering rate limit is 3× the momentum scattering rate low energy limit: $1/\tau_q^{POP} = 2\omega_{LO}\alpha_{LO}N_{LO}$. Both rates decrease with increasing carrier energy: in this aspect POP scattering is similar to Coulomb impurity scattering; both have charge-based origin. Note that the quantum scattering rate is larger than the momentum scattering rate at all energies. The plots are for $z = \frac{\hbar\omega_{LO}}{k_bT} = 1$. Equations 23.80 and 23.81 state that the relative importance of the emission rate over the absorption rate is enhanced as e^z, implying an enhanced emission rate in semiconductors with large $\hbar\omega_{LO}$, which enables simplifications in the evaluation of the carrier mobilty.

Because of its inelastic nature, and since the POP rates do not have a simple energy dependence, there is no general analytical expression for the resulting carrier mobility. Nevertheless, for optical phonon energies that satisfy $\hbar\omega_{LO} >> k_bT$, a reasonably accurate approximate expression results from the BTE by assuming a much higher emission rate than the absorption rate. The resulting carrier mobility corresponds to the low-energy limit of the POP quantum scattering time:

$$\mu_{3d}^{POP} \approx \frac{q\tau_q^{POP}(E_{\mathbf{k}} \to 0)}{m^\star} = \frac{q}{2m^\star\alpha_{LO}\omega_{LO}N_{LO}} \approx \frac{q \cdot \exp[-\frac{\hbar\omega_{LO}}{k_bT}]}{2m^\star\alpha_{LO}\omega_{LO}}. \tag{23.82}$$

With the expressions for all 4 phonon mode scattering mechanisms (ADP, PZ, ODP, and POP) at hand, it is instructive to gauge their relative importance. Fig. 23.36 (a) shows the scattering rates of all four phonon modes for GaAs, and Fig. 23.36 (b) shows the numerically calculated electron mobility as a function of temperature. The POP scattering rate dominates both at low and high energies. POP and ODP[29] emission have a threshold at $E_{\mathbf{k}} = \hbar\omega_{LO}$. Because POP and PZ scattering originate from charge, they have a decreasing trend with energy unlike ADP and ODP mechanisms. The calculated mobilities near room temperature (close to experimental values) are limited mainly by POP scattering. At the lowest temperatures the mobility is limited intrinsically by acoustic mode scattering, but this is only observable in ultraclean samples in which other scattering mechanisms are absent.

All phonon modes, 2D carriers: For quantum confined carriers whose motion is restricted to lower dimensions (such as in quantum wells or wires), the spirit of the above formulation for phonon scattering carries over, but the matrix elements must now explicitly account for the modified wavefunctions of electrons[30]. We now re-revaluate the matrix elements for 2D electrons. Because we consider bulk phonon modes, the scattering potentials of the four modes remain the same as in Table 23.6. Consider quantum confinement of electrons in the z-direction: the spatial vector is written as $\mathbf{r} = \mathbf{r}_{\parallel} + z\hat{z} = \rho\hat{\rho} + z\hat{z}$, where $\mathbf{r}_{\parallel} = \rho\hat{\rho}$ is the 2D vector with $\hat{\rho}$ a unit vector in the $x-y$ plane of area L^2, and $\hat{z}$ the unit vector in the z-direction. The 3D electron wavevector is split into $\mathbf{k} = \mathbf{k}_{\parallel} + k_z\hat{z} = k_{\parallel}\hat{\rho} + k_z\hat{z}$, and a similar decomposition is made for the phonon wavevector: $\vec{\beta} = \vec{\beta}_{\parallel} + \beta_z\hat{z} = \beta_{\parallel}\hat{\rho} + \beta_z\hat{z}$. Because potential gradients exist in the z-direction to confine the electrons, but no gradients exist in the $x-y$ plane, scattering events with phonons will conserve the net momentum only in the $x-y$ plane, and *not* in the z-direction. Taking these facts into account, the matrix element for scattering of 2D electrons by phonons is then

$$\langle f|W_0(\mathbf{r})|i\rangle = \int d^3r \cdot \psi_f^{\star}(\mathbf{r}) \cdot W_0(\mathbf{r}) \cdot \psi_i(\mathbf{r})$$

$$= \int d^3r [\frac{e^{-i\mathbf{k}'_{\parallel}\cdot\mathbf{r}_{\parallel}}}{\sqrt{L^2}} \chi_f(z) u^{\star}_{\mathbf{k}'}(\mathbf{r})][K_\beta \cdot u_\beta \cdot e^{\pm i\vec{\beta}_{\parallel}\cdot\mathbf{r}_{\parallel}} \cdot e^{\pm i\beta_z z}][\frac{e^{i\mathbf{k}_{\parallel}\cdot\mathbf{r}_{\parallel}}}{\sqrt{L^2}} \chi_i(z) u_{\mathbf{k}}(\mathbf{r})]$$

$$= K_\beta u_\beta I_{\mathbf{k},\mathbf{k}'} \cdot \underbrace{\int dz \cdot \chi_f(z) \cdot e^{\pm i\beta_z z} \cdot \chi_i(z)}_{F(\beta_z)} \cdot \underbrace{\int \frac{d^2r}{L^2} e^{i(\mathbf{k}_{\parallel} - \mathbf{k}'_{\parallel} \pm \vec{\beta}_{\parallel})\cdot\mathbf{r}_{\parallel}}}_{\delta_{\mathbf{k}'_{\parallel},\mathbf{k}_{\parallel}\pm\vec{\beta}_{\parallel}}}$$

$$\Longrightarrow \langle f|W_0(\mathbf{r})|i\rangle \approx K_\beta \cdot u_\beta \cdot F(\beta_z) \cdot \delta_{\mathbf{k}'_{\parallel},\mathbf{k}_{\parallel}\pm\vec{\beta}_{\parallel}}$$

$$\Longrightarrow \boxed{|\langle f|W_0(\mathbf{r})|i\rangle|^2 = |K_\beta|^2 \cdot |u_\beta|^2 \cdot |F(\beta_z)|^2 \cdot \delta_{\mathbf{k}'_{\parallel},\mathbf{k}_{\parallel}\pm\vec{\beta}_{\parallel}}.}$$

Comparing this result to Equation 23.70 highlights that the major difference from the 3D phonon scattering case is the appearance of the form factor $|F(\beta_z)|^2$ and the requirement of in-plane rather than 3D momentum conservation. The rest of the procedure involves the direct evaluation of the scattering rates, for which the quantum confined envelope functions $\chi(z)$ in the z-direction must be known. We proceed as far as possible now with the general $\chi(z)$ first, and then discuss a few particular choices.

---

[29] Note that strictly speaking the ODP mechanism is forbidden for carriers in the direct Γ-valleys, so its inclusion for GaAs is purely for comparison. Similarly, POP and PZ scattering are absent in elemental semiconductors such as Si, Ge, and diamond. The next section discusses these matters in more detail.

[30] To be precise, in addition to the electron states, the phonon modes and their spectra are also modified in semiconductor heterostructures because of quantum confinement. The effect of confined phonons can be treated on the same footing as that of the electrons. Because experimental evidence suggests the phonon confinement effect not to be substantial in most semiconductor heterostructures, and because the treatment simplifies considerably by considering bulk phonon modes, we cover this (bulk phonon + quantum confined electron) situation in the discussion here.

Acoustic deformation potential, 2D carriers: Consider ADP scattering first, for which $K_\beta^2 = \beta^2 D_a^2 = (\beta_\parallel^2 + \beta_z^2) D_a^2$. For the momentum scattering rate for 2D carriers we have the following explicit derivation:

$$\frac{1}{\tau_{m,abs}^{ADP}(\mathbf{k}_\parallel)} = \sum_\beta \sum_{\mathbf{k}'_\parallel} (1 - \frac{\mathbf{k}'_\parallel \cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \cdot S_{abs}(\mathbf{k}, \mathbf{k}')$$

$$= \sum_\beta \sum_{\mathbf{k}'_\parallel} (1 - \frac{\mathbf{k}'_\parallel \cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \frac{2\pi}{\hbar} |K_\beta|^2 |u_\beta^{abs}|^2 |F(\beta_z)|^2 \delta_{\mathbf{k}'_\parallel, \mathbf{k}_\parallel \pm \vec{\beta}_\parallel} \delta[E_f - (E_i + \hbar\omega_\beta)]$$

$$= \frac{2\pi}{\hbar} \sum_\beta (1 - \frac{(\mathbf{k}_\parallel + \vec{\beta}_\parallel) \cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \beta^2 D_a^2 \frac{\hbar N_\beta}{2\rho L^3 \omega_\beta} |F(\beta_z)|^2 \delta[E_f - (E_i + \hbar\omega_\beta)]$$

$$= \frac{\pi D_a^2}{\rho L^3} \sum_\beta |F(\beta_z)|^2 \cdot (-\frac{\beta_\parallel}{k_\parallel} \cos\alpha) \cdot \beta^2 \cdot \frac{k_b T}{\hbar \omega_\beta^2} \times$$

$$\delta[\Big(\mathcal{E}_f + \frac{\hbar^2 |\mathbf{k}_\parallel + \vec{\beta}_\parallel|^2}{2m^\star}\Big) - \Big(\mathcal{E}_i + \frac{\hbar^2 |\mathbf{k}_\parallel|^2}{2m^\star} + \hbar v_s \beta\Big)]$$

$$= \frac{\pi D_a^2 k_b T}{\hbar \rho v_s^2 L^3 k_\parallel} \sum_\beta |F(\beta_z)|^2 (-\beta_\parallel \cos\alpha) \frac{2m^\star}{\hbar^2} \delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - 2\beta_\parallel k_s + k_{if}^2]$$

$$= \frac{2\pi m^\star D_a^2 k_b T}{\hbar^3 \rho v_s^2 L^3 k_\parallel} \sum_\beta |F(\beta_z)|^2 (-\beta_\parallel \cos\alpha) \cdot \delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - 2\beta_\parallel k_s + k_{if}^2]$$

$$= \frac{2\pi m^\star D_a^2 k_b T}{\hbar^3 \rho v_s^2 L^3 k_\parallel} \int \frac{d^2\beta_\parallel (-\beta_\parallel \cos\alpha)}{\frac{(2\pi)^2}{L^2}} \cdot \underbrace{\delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - 2\beta_\parallel k_s + k_{if}^2]}_{=\delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha] \text{ for } \mathcal{E}_f = \mathcal{E}_f \text{ and } k_s << k_\parallel} \times$$

$$\frac{1}{\frac{2\pi}{L}} \cdot \underbrace{\int_{-\infty}^{+\infty} d\beta_z |F(\beta_z)|^2}_{\mathcal{K}}$$

$$= \frac{m^\star D_a^2 k_b T \mathcal{K}}{(2\pi)^2 \hbar^3 \rho v_s^2 k_\parallel} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} d\alpha (-\cos\alpha) \int_0^{2k_\parallel} d\beta_\parallel \cdot \beta_\parallel^2 \cdot \underbrace{\delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha]}_{= \frac{\delta[\beta_\parallel] + \delta[\beta_\parallel + 2k_\parallel \cos\alpha]}{-2k_\parallel \cos\alpha}}$$

$$= \frac{m^\star D_a^2 k_b T \mathcal{K}}{(2\pi)^2 \hbar^3 \rho v_s^2 k_\parallel} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} d\alpha (-\cos\alpha)(-2k_\parallel \cos\alpha) = \frac{m^\star D_a^2 k_b T}{(2\pi)^2 \hbar^3 \rho v_s^2 k_\parallel} \cdot K \cdot \pi k_\parallel$$

$$\implies \frac{1}{\tau_{m,abs}^{ADP}(\mathbf{k}_\parallel)} = \frac{m^\star D_a^2 k_b T}{4\pi \hbar^3 \rho v_s^2} \cdot \mathcal{K} \text{ and } \frac{1}{\tau_{m,em}^{ADP}(\mathbf{k}_\parallel)} \approx \frac{1}{\tau_{m,bas}^{ADP}(\mathbf{k}_\parallel)}$$

$$\implies \boxed{\frac{1}{\tau_{m,2D}^{ADP}(\mathbf{k}_\parallel)} \approx \frac{2}{\tau_{m,abs}^{ADP}(\mathbf{k}_\parallel)} = \frac{m^\star D_a^2 k_b T}{2\pi \hbar^3 \rho v_s^2} \cdot \mathcal{K}.} \tag{23.83}$$

Note that the final 2D ADP scattering rate is similar to the 3D case upon the replacement $k \to \mathcal{K}/2$, and is *independent* of $\mathbf{k}_{\|}$. Since $\tau_m = \tau_0$ with no energy dependence, using $(d = 2, r = 0)$ in Equation 23.35 results in an exact expression for the 2D ADP scattering limited mobility:

$$\mu_{2d}^{ADP} = \frac{q\tau_0}{m^\star}\cdot\frac{\Gamma(2+r)}{\Gamma(2)}\cdot\frac{F_r(\eta)}{F_0(\eta)} \implies \boxed{\mu_{2d}^{ADP} = \frac{2\pi e\hbar^3\rho v_s^2}{(m^\star)^2 D_a^2 \mathcal{K} k_b T}.} \tag{23.84}$$

Though $\mathcal{K}$ should be found numerically in general, analytical values can be looked up from Table 23.2. For the particular case of an infinite well of thickness L_z, the value is $\mathcal{K} = \frac{3\pi}{L_z}$. For the Fang–Howard–Stern variational wavefunction of parameter b it is $\mathcal{K} = \frac{3\pi b}{8}$. Note that unlike the $T^{-\frac{3}{2}}$ dependence of the mobility limited by ADP scattering in 3D, in 2D it goes as T^{-1}. Also, the ADP limited mobility is smaller for thinner wells (smaller L_z), or for larger 2D carrier densities (via the FHS parameter b).

Piezoelectric scattering, 2D carriers: The 2D PZ scattering rate is

$$\frac{1}{\tau_{m,abs}^{PZ}(\mathbf{k}_{\|})} = \sum_{\beta}\sum_{\mathbf{k}'_{\|}}\left(1 - \frac{\mathbf{k}'_{\|}\cdot\mathbf{k}_{\|}}{|\mathbf{k}_{\|}|^2}\right)\cdot S_{abs}(\mathbf{k},\mathbf{k}')$$

$$= \frac{2\pi}{\hbar}\sum_{\beta}\left(1 - \frac{(\mathbf{k}_{\|}+\vec{\beta}_{\|})\cdot\mathbf{k}_{\|}}{|\mathbf{k}_{\|}|^2}\right)\cdot\frac{q^2 e_{pz}^2}{\epsilon_{dc}^2}\cdot\frac{\hbar N_\beta}{2\rho L^3\omega_\beta}\cdot|F(\beta_z)|^2 \times$$

$$\frac{2m^\star}{\hbar^2}\delta[\beta_{\|}^2 + 2\beta_{\|}k_{\|}\cos\alpha - 2\beta_{\|}k_s + k_{if}^2]$$

$$= \frac{2\pi m^\star q^2 e_{pz}^2 k_b T}{\hbar^3\rho v_s^2\epsilon_{dc}^2 L^3}\sum_{\beta}\frac{|F(\beta_z)|^2}{\beta^2}\left(-\frac{\beta_{\|}\cos\alpha}{k_{\|}}\right)\cdot\delta[\beta_{\|}^2 + 2\beta_{\|}k_{\|}\cos\alpha]$$

$$= \frac{m^\star q^2 e_{pz}^2 k_b T}{(2\pi)^2\hbar^3\rho v_s^2\epsilon_{dc}^2 k_{\|}} \times$$

$$\int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} d\alpha(-\cos\alpha)\cdot\int_0^{2k_{\|}} d\beta_{\|}\cdot\beta_{\|}^2\cdot\delta[\beta_{\|}^2 + 2\beta_{\|}k_{\|}\cos\alpha]\cdot\underbrace{\int_{-\infty}^{+\infty} d\beta_z\frac{|F(\beta_z)|^2}{\beta_{\|}^2+\beta_z^2}}_{\mathcal{G}(\beta_{\|})}$$

$$\frac{1}{\tau_{m,em}^{PZ}(\mathbf{k}_{\|})} = \frac{m^\star q^2 e_{pz}^2 k_b T}{\pi^2\hbar^3\rho v_s^2\epsilon_{dc}^2}\cdot\int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} d\alpha\cdot(\cos\alpha)^2\cdot\mathcal{G}(-2k_{\|}\cos\alpha)$$

$$\frac{1}{\tau_{m,em}^{PZ}(\mathbf{k}_{\|})} \approx \frac{1}{\tau_{m,abs}^{PZ}(\mathbf{k}_{\|})} \implies \frac{1}{\tau_{m,2D}^{PZ}(\mathbf{k}_{\|})} \approx \frac{2}{\tau_{m,abs}^{PZ}(\mathbf{k}_{\|})}$$

$$\implies \boxed{\frac{1}{\tau_{m,2D}^{PZ}(\mathbf{k}_{\|})} = \frac{2m^\star q^2 e_{pz}^2 k_b T}{\pi^2\hbar^3\rho v_s^2\epsilon_{dc}^2}\cdot\int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} d\alpha\cdot(\cos\alpha)^2\cdot\mathcal{G}(-2k_{\|}\cos\alpha).} \tag{23.85}$$

The function $\mathcal{G}(...)$ has dimensions of length, the values of which for the three cases of the δ-well, the ∞-well and the Fang–Howard–Stern variational wavefunctions are given in Table 23.2. For general shapes of envelope functions $\chi(z)$, it should be evaluated numerically.

Optical deformation potential, 2D carriers: We get for 2D ODP[31]

$$
\begin{aligned}
\frac{1}{\tau_{m,abs}^{ODP}(\mathbf{k}_\parallel)} &= \sum_\beta \sum_{\mathbf{k}'_\parallel} (1 - \frac{\mathbf{k}'_\parallel \cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \cdot S_{abs}(\mathbf{k}, \mathbf{k}') \\
&= \frac{2\pi}{\hbar} \sum_\beta (1 - \frac{(\mathbf{k}_\parallel + \vec{\beta}_\parallel) \cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \cdot D_{op}^2 \cdot \frac{\hbar N_{LO}}{2\rho L^3 \omega_{LO}} \cdot |F(\beta_z)|^2 \times \\
&\qquad \frac{2m^\star}{\hbar^2} \delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - k_{LO}^2 + k_{if}^2] \\
&= \frac{2\pi m^\star D_{op}^2 N_{LO}}{\hbar^2 \rho \omega_{LO} L^3} \sum_\beta |F(\beta_z)|^2 \cdot (-\frac{\beta_\parallel}{k_\parallel} \cos\alpha) \cdot \delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - k_{LO}^2 + k_{if}^2] \\
&= \frac{m^\star D_{op}^2 N_{LO}}{4\pi^2 \hbar^2 \rho \omega_{LO} k_\parallel} \cdot \underbrace{\int_{-\infty}^{+\infty} d\beta_z |F(\beta_z)|^2}_{\mathcal{K}} \times \\
&\qquad \int_0^{2\pi} d\alpha(-\cos\alpha) \cdot \int_{\beta_{\parallel,min}^{abs}}^{\beta_{\parallel,max}^{abs}} d\beta_\parallel \cdot \beta_\parallel^2 \cdot \underbrace{\delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - k_{LO}^2 + k_{if}^2]}_{= \frac{\delta[\beta_\parallel - \beta_+^{abs}] + \delta[\beta_\parallel - \beta_-^{abs}]}{\beta_+^{abs} - \beta_-^{abs}} \text{ for } k_{if=0}} \\
&= \frac{m^\star D_{op}^2 N_{LO}}{8\pi^2 \hbar^2 \rho \omega_{LO}} \cdot \mathcal{K} \cdot \underbrace{\int_0^{2\pi} d\alpha(-\cos\alpha) \cdot \frac{(\sqrt{\cos^2\alpha + (\frac{k_{LO}}{k_\parallel})^2} - \cos\alpha)^2}{\sqrt{\cos^2\alpha + (\frac{k_{LO}}{k_\parallel})^2}}}_{=2\pi} \\
\implies \frac{1}{\tau_{m,abs}^{ODP}(\mathbf{k}_\parallel)} &= \frac{m^\star D_{op}^2 N_{LO}}{4\pi \hbar^2 \rho \omega_{LO}} \cdot \mathcal{K} \\
\frac{1}{\tau_{m,em}^{ODP}(\mathbf{k}_\parallel)} &= \sum_\beta \sum_{\mathbf{k}'_\parallel} (1 + \frac{\mathbf{k}'_\parallel \cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \cdot S_{em}(\mathbf{k}, \mathbf{k}') \\
&= \frac{m^\star D_{op}^2 N_{LO}}{4\pi^2 \hbar^2 \rho \omega_{LO} k_\parallel} \cdot \underbrace{\int_{-\infty}^{+\infty} d\beta_z |F(\beta_z)|^2}_{\mathcal{K}} \times \\
&\qquad \int_{-\alpha_{max}}^{+\alpha_{max}} d\alpha(+\cos\alpha) \cdot \int_{\beta_{\parallel,min}^{em}}^{\beta_{\parallel,max}^{em}} d\beta_\parallel \cdot \beta_\parallel^2 \cdot \underbrace{\delta[\beta_\parallel^2 - 2\beta_\parallel k_\parallel \cos\alpha + k_{LO}^2]}_{= \frac{\delta[\beta_\parallel - \beta_+^{em}] + \delta[\beta_\parallel - \beta_-^{em}]}{\beta_+^{em} - \beta_-^{em}} \text{ for } k_{if=0}} \\
&= \frac{m^\star D_{op}^2 N_{LO}}{4\pi^2 \hbar^2 \rho \omega_{LO} k_\parallel} \cdot \mathcal{K} \cdot \underbrace{\int_{-\alpha_{max}}^{+\alpha_{max}} d\alpha(+\cos\alpha) \cdot \frac{(\beta_+^{em})^2 + (\beta_-^{em})^2}{\beta_+^{em} - \beta_-^{em}}}_{=\pi k_\parallel} \\
\implies \frac{1}{\tau_{m,em}^{ODP}(\mathbf{k}_\parallel)} &= \frac{m^\star D_{op}^2 (N_{LO} + 1)}{4\pi \hbar^2 \rho \omega_{LO}} \cdot \mathcal{K} \cdot \Theta(E_{\mathbf{k}_\parallel} - \hbar\omega_{LO}).
\end{aligned}
\tag{23.86}
$$

---

[31] The following relations are useful for the argument of the Dirac delta function for optical phonon scattering. The roots for absorption are $\beta_+^{abs} = k_\parallel(\sqrt{\cos^2\alpha + (\frac{k_{LO}}{k_\parallel})^2} - \cos\alpha)$ and $\beta_-^{abs} = k_\parallel(-\sqrt{\cos^2\alpha + (\frac{k_{LO}}{k_\parallel})^2} - \cos\alpha) < 0$. Since $\beta \geq 0$ is a radial coordinate, β_-^{abs} is not allowed. Note that $\beta_+^{abs} - \beta_-^{abs} = 2k_\parallel \sqrt{\cos^2\alpha + (\frac{k_{LO}}{k_\parallel})^2}$. Similarly, for emission processes, the roots of the Dirac delta function are $\beta_+^{em} = k_\parallel(\cos\alpha + \sqrt{\cos^2\alpha - (\frac{k_{LO}}{k_\parallel})^2})$ and $\beta_-^{em} = k_\parallel(\cos\alpha - \sqrt{\cos^2\alpha - (\frac{k_{LO}}{k_\parallel})^2})$, implying $\beta_+^{em} - \beta_-^{em} = 2k_\parallel \sqrt{\cos^2\alpha - (\frac{k_{LO}}{k_\parallel})^2}$. The range of angles for emission is restricted from $-\alpha_{max} \leq \alpha \leq \alpha_{max}$, where $\cos\alpha_{max} = \frac{k_{LO}}{k_\parallel}$. These are used in the derivation of Equation 23.86.

The combined 2D ODP momentum relaxation rate is thus[32]

$$\boxed{\frac{1}{\tau_{m,2D}^{ODP}(\mathbf{k}_\parallel)} = \frac{m^\star D_{op}^2 \mathcal{K} N_{LO}}{4\pi\hbar^2\rho\omega_{LO}}\Big[1 + e^{\frac{\hbar\omega_{LO}}{k_b T}}\Theta(E_{\mathbf{k}_\parallel} - \hbar\omega_{LO})\Big].} \tag{23.87}$$

Polar optical phonon, 2D carriers: Finally, consider POP scattering in 2D. The same angular restrictions on α and wavevector restrictions on $\beta_\parallel$ apply as for ODP to give scattering rates

$$\begin{aligned}
\frac{1}{\tau_{m,abs}^{POP}(\mathbf{k}_\parallel)} &= \sum_{\beta}\sum_{\mathbf{k}'_\parallel}(1 - \frac{\mathbf{k}'_\parallel \cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \cdot S_{abs}(\mathbf{k}, \mathbf{k}') \\
= \frac{2\pi}{\hbar}\sum_{\beta}(1 - &\frac{(\mathbf{k}_\parallel + \vec{\beta}_\parallel)\cdot \mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2}) \cdot \frac{C_0^2}{\beta^2} \cdot \frac{\hbar N_{LO}}{2\rho L^3 \omega_{LO}} \cdot |F(\beta_z)|^2 \times \\
&\frac{2m^\star}{\hbar^2}\delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - k_{LO}^2 + k_{if}^2] \\
= \frac{2\pi m^\star C_0^2 N_{LO}}{\hbar^2 \rho \omega_{LO} L^3}\sum_{\beta}\frac{|F(\beta_z)|^2}{\beta^2}(-&\frac{\beta_\parallel}{k_\parallel}\cos\alpha)\delta[\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - k_{LO}^2 + k_{if}^2] \\
&= \frac{m^\star C_0^2 N_{LO}}{4\pi^2\hbar^2\rho\omega_{LO}k_\parallel} \times \\
\int_0^{2\pi} d\alpha(-\cos\alpha)\int_{\beta_{\parallel,min}^{abs}}^{\beta_{\parallel,max}^{abs}} d\beta_\parallel \cdot \beta_\parallel^2 &\delta[\underbrace{\beta_\parallel^2 + 2\beta_\parallel k_\parallel \cos\alpha - k_{LO}^2}_{=(\beta_\parallel - \beta_+^{abs})(\beta_\parallel - \beta_-^{abs})}]\underbrace{\int_{-\infty}^{+\infty} d\beta_z \frac{|F(\beta_z)|^2}{\beta_\parallel^2 + \beta_z^2}}_{\mathcal{G}(\beta_\parallel)} \\
&= \frac{m^\star C_0^2 N_{LO}}{4\pi^2\hbar^2\rho\omega_{LO}k_\parallel} \times \\
\int_0^{2\pi} d\alpha(-\cos\alpha)\cdot\int_{\beta_{\parallel,min}^{abs}}^{\beta_{\parallel,max}^{abs}} d\beta_\parallel \cdot \beta_\parallel^2 \cdot &\mathcal{G}(\beta_\parallel) \cdot [\frac{\delta[\beta_\parallel - \beta_+^{abs}] + \delta[\beta_\parallel - \beta_-^{abs}]}{\beta_+^{abs} - \beta_-^{abs}}] \\
&= \frac{m^\star C_0^2 N_{LO}}{4\pi^2\hbar^2\rho\omega_{LO}k_\parallel} \times \int_0^{2\pi} d\alpha(-\cos\alpha)\cdot\frac{(\beta_+^{abs})^2 \cdot \mathcal{G}(\beta_+^{abs})}{\beta_+^{abs} - \beta_-^{abs}} \\
= \frac{\omega_{LO}\alpha_{LO}N_{LO}}{2\pi}&\underbrace{[\, k_{LO}\int_0^{2\pi} d\alpha(-\cos\alpha)\frac{(\sqrt{\cos^2\alpha + (\frac{k_{LO}}{k_\parallel})^2} - \cos\alpha)^2 \cdot \mathcal{G}(\beta_+^{abs})}{\sqrt{\cos^2\alpha + (\frac{k_{LO}}{k_\parallel})^2}}\,]}_{I_{2d}^{abs}(k_\parallel)} \\
\implies \frac{1}{\tau_{m,abs}^{POP}(\mathbf{k}_\parallel)} &= \frac{\omega_{LO}}{2\pi}\cdot\alpha_{LO}\cdot N_{LO}\cdot I_{2d}^{abs}(k_\parallel).
\end{aligned} \tag{23.88}$$

The functional form of $\mathcal{G}(...)$ must be known to evaluate $I_{2d}^{abs}(k_\parallel)$.

---

[32] The mobility is obtained similar to the 3D ODP scattering case.

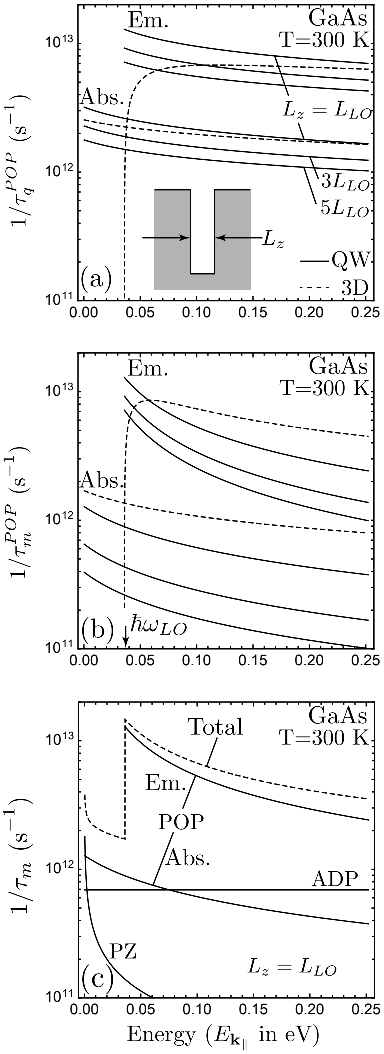


Fig. 23.37 (a) Quantum scattering rates and (b) momentum scattering rates of 2D carriers by polar optical phonons in GaAs quantum wells of thickness L_z at 300 K. The dashed lines are the corresponding 3D scattering rates. Here $L_0 = 1/k_{LO} = \sqrt{\hbar/2m^\star\omega_{LO}}$ is a characteristic length for POP scattering, about 4 nm for GaAs. (c) A comparison of the relative strengths of the four phonon modes via their 300 K momentum scattering rates.

Using a similar procedure, for optical phonon emission we get

$$\frac{1}{\tau^{POP}_{m,em}(\mathbf{k}_\parallel)} = \sum_{\beta}\sum_{\mathbf{k}'_\parallel}(1-\frac{\mathbf{k}'_\parallel\cdot\mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2})\cdot S_{em}(\mathbf{k},\mathbf{k}')$$

$$= \frac{2\pi}{\hbar}\sum_{\beta}(1-\frac{(\mathbf{k}_\parallel-\vec{\beta}_\parallel)\cdot\mathbf{k}_\parallel}{|\mathbf{k}_\parallel|^2})\cdot\frac{C_0^2}{\beta^2}\cdot\frac{\hbar(N_{LO}+1)}{2\rho L^3\omega_{LO}}\cdot|F(\beta_z)|^2\times$$

$$\frac{2m^\star}{\hbar^2}\delta[\beta_\parallel^2-2\beta_\parallel k_\parallel\cos\alpha+k_{LO}^2]$$

$$= \frac{m^\star C_0^2(N_{LO}+1)}{4\pi^2\hbar^2\rho\omega_{LO}k_\parallel}\times$$

$$\int_{-\alpha_{max}}^{+\alpha_{max}} d\alpha(\cos\alpha)\int_{\beta^{em}_{\parallel,min}}^{\beta^{em}_{\parallel,max}} d\beta_\parallel\cdot\beta_\parallel^2\delta\underbrace{[\beta_\parallel^2-2\beta_\parallel k_\parallel\cos\alpha+k_{LO}^2]}_{=(\beta_\parallel-\beta_+^{em})(\beta_\parallel-\beta_-^{em})}\underbrace{\int_{-\infty}^{+\infty}d\beta_z\frac{|F(\beta_z)|^2}{\beta_\parallel^2+\beta_z^2}}_{\mathcal{G}(\beta_\parallel)}$$

$$= \frac{m^\star C_0^2(N_{LO}+1)}{4\pi^2\hbar^2\rho\omega_{LO}k_\parallel}\times\int_{-\alpha_{max}}^{+\alpha_{max}} d\alpha(\cos\alpha)\cdot\Big[\frac{(\beta_+^{em})^2\cdot\mathcal{G}(\beta_+^{em})+(\beta_-^{em})^2\cdot\mathcal{G}(\beta_-^{em})}{\beta_+^{em}-\beta_-^{em}}\Big]$$

$$= \frac{\omega_{LO}\alpha_{LO}(N_{LO}+1)}{2\pi}\times$$

$$\Big[\underbrace{k_{LO}\int_{-\alpha_{max}}^{+\alpha_{max}} d\alpha(\cos\alpha)\frac{(\cos\alpha+\sqrt{\cos^2\alpha-(\frac{k_{LO}}{k_\parallel})^2})^2\cdot\mathcal{G}(\beta_+^{em})}{\sqrt{\cos^2\alpha-(\frac{k_{LO}}{k_\parallel})^2}}}_{I^{em}_{2d,+}(k_\parallel)}+$$

$$\underbrace{k_{LO}\int_{-\alpha_{max}}^{+\alpha_{max}} d\alpha(\cos\alpha)\frac{(\cos\alpha-\sqrt{\cos^2\alpha-(\frac{k_{LO}}{k_\parallel})^2})^2\cdot\mathcal{G}(\beta_-^{em})}{\sqrt{\cos^2\alpha-(\frac{k_{LO}}{k_\parallel})^2}}}_{I^{em}_{2d,-}(k_\parallel)}\Big]$$

$$\implies \frac{1}{\tau^{POP}_{m,em}(\mathbf{k}_\parallel)} = \frac{\omega_{LO}\alpha_{LO}(N_{LO}+1)}{2\pi}\big[I^{em}_{2d,+}(k_\parallel)+I^{em}_{2d,-}(k_\parallel)\big]\Theta[E_{\mathbf{k}_\parallel}-\hbar\omega_{LO}]. \tag{23.89}$$

Combining the absorption and emission rates gives the total momentum scattering rate due to polar optical phonons for 2D carriers:

$$\frac{1}{\tau^{POP}_{m,2D}(\mathbf{k}_\parallel)} = \frac{\omega_{LO}\alpha_{LO}N_{LO}}{2\pi}\times$$

$$\Big[\underbrace{I^{abs}_{2d}(k_\parallel)}_{abs.}+\underbrace{e^{\frac{\hbar\omega_{LO}}{k_bT}}\big[I^{em}_{2d,+}(k_\parallel)+I^{em}_{2d,-}(k_\parallel)\big]\cdot\Theta(E_{\mathbf{k}_\parallel}-\hbar\omega_{LO})}_{em.}\Big], \tag{23.90}$$

which is an exact (and powerful) relation that can be used for 2D electron and hole gases confined in quantum wells with arbitrary $\chi(z)$. For most cases, the rates must be evaluated numerically. For three specific cases: the δ-confinement, the infinitely deep well, and for the FHS cases, Table 23.2 provides the analytical forms of $\mathcal{G}(x)$ needed for calculating the POP scattering rates in 2D.

The quantum relaxation rates due to POP in 2D given by

$$\frac{1}{\tau_{q,abs}^{POP}(\mathbf{k}_{\|})} = \frac{\omega_{LO}\alpha_{LO}N_{LO}k_{LO}}{\pi} \int_0^{2\pi} d\alpha \frac{\beta_+^{abs}\mathcal{G}(\beta_+^{abs})}{\beta_+^{abs} - \beta_-^{abs}}$$

$$\frac{1}{\tau_{q,em}^{POP}(\mathbf{k}_{\|})} = \frac{\omega_{LO}\alpha_{LO}(N_{LO}+1)k_{LO}}{\pi} \int_{-\alpha_{max}}^{+\alpha_{max}} d\alpha \Big[\frac{\beta_+^{em}\mathcal{G}(\beta_+^{em}) + \beta_-^{em}\mathcal{G}(\beta_-^{em})}{\beta_+^{em} - \beta_-^{em}}\Big], \tag{23.91}$$

are, similar to the 3D case of POP scattering, larger than the momentum relaxation rates. Figures 23.37 (a) and (b) compare the momentum and quantum relaxation rates due to POP phonons for 2D electron gases in GaAs quantum wells of varying thickness, highlighting the effect of quantum confinement. The corresponding 3D scattering rates are also shown in the same plots for comparison. Note that the scattering rate decreases for thicker wells, and that the emission rate just above the threshold $E_{\mathbf{k}_{\|}} \geq \hbar\omega_{LO}$ is larger in 2D than in 3D.

The resulting 2D carrier mobility limited by POP scattering must be evaluated numerically using Equation 23.90 with Equation 23.35, and no simple expression exists for the general case. But just as in the case of POP scattering in 3D, the following reasonably accurate expression is derived for the case when $\hbar\omega_{LO} >> k_b T$:

$$\mu_{2d}^{POP} \approx \frac{e}{m^\star} \cdot \frac{1}{\omega_{LO}\alpha_{LO}N_{LO}} \cdot \frac{1 + \frac{1-e^{-y}}{y}}{2\pi P_1(k_{LO})}, \tag{23.92}$$

where $y = \frac{\pi\hbar^2 n_{2d}}{m^\star k_b T}$ and $P_1(x)$ is the form-factor in Table 23.2. POP scattering limited mobility of 2D electron gases in compound semiconductor heterostrutures such as in GaAs and GaN are explained reasonably well with this approximate form.

A comparison of the relative strengths of scattering in 2D of all four phonon modes for an AlGaAs/GaAs system is shown in Fig. 23.37 (c), which should be compared to Fig. 23.36 (a) for the corresponding 3D situation. Since the ADP scattering rate is proportional to the density of states, it increases as $\sqrt{E_{\mathbf{k}}}$ in 3D but is constant in 2D.

Finally, Fig. 23.38 shows the resulting phonon limited mobilities for 2D electron gases in AlGaAs/GaAs and AlGaN/GaN heterostructures. Note that for both technologically important semiconductors the mobilities for the conditions shown are limited by POP scattering at room-temperature, and by acoustic phonon scattering at low-temperatures. It is emphasized that these are the estimates of the *intrinsic* phonon limited mobilities; defects and disorder will only lower them further, which is the subject of discussion of the next section.

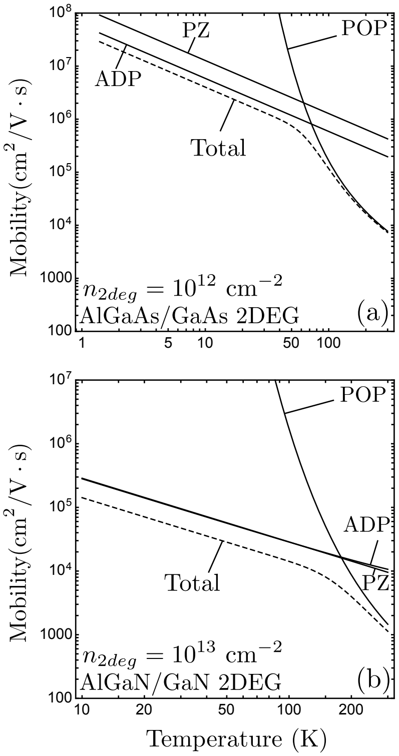


Fig. 23.38 Phonon limited 2DEG mobilities in (a) AlGaAs/GaAs and (b) AlGaN/GaN hetersotructures. Experimental values at 300 K are ∼ 8000 cm^2/V·s for AlGaAs/GaAs and ∼ 1500 cm^2/V·s for AlGaN/GaN for these densities, similar to the calculated values. Note that Equation 23.92 is used to calculate the POP limited 2DEG mobilities.

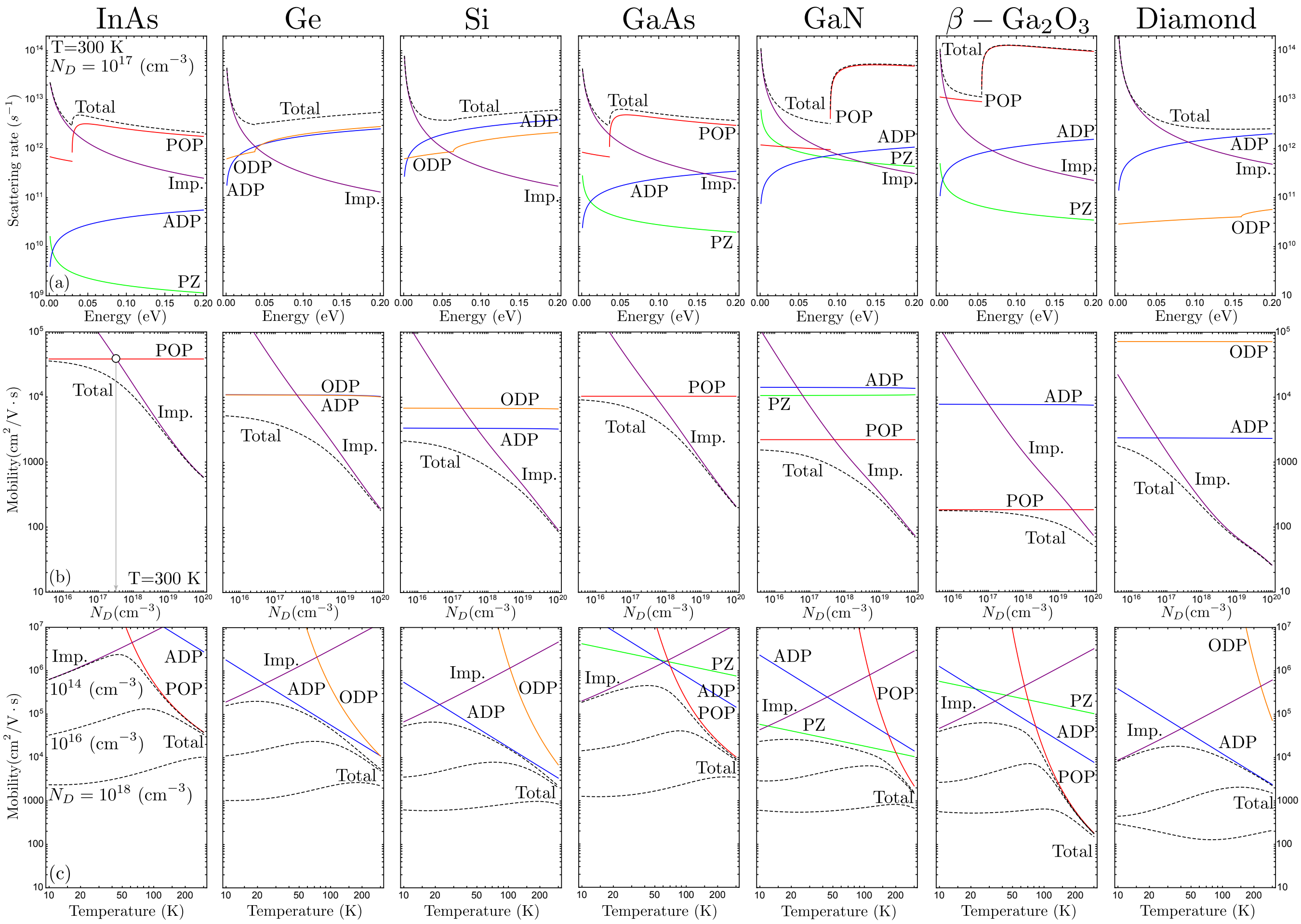


Fig. 23.39 Calculated 3D scattering rates and electron mobilities in various semiconductors. Horizontal panel (a) compares the 300 K scattering rates at a fixed doping density, panel (b) the mobility vs. the doping density at 300 K, and panel (c) the mobility vs. temperature for three doping densities.

23.10 Experimental mobilities

Fig. 23.39 shows the effect of doping on the mobilities for various semiconductors[33]. Note that unlike the compound semiconductors that have the conduction band minima at the Γ-point, the indirect gap elemental semiconductors Ge, Si and diamond have no piezoelectric or polar optical phonon scattering, but have optical deformation potential scattering which is absent for Γ-valley conduction minima. These should be compared with the general experimental trend that was mentioned in Fig. 23.1 in the beginning of this chapter. Most experimental data, though not explicitly plotted on this figure, fall reasonably close to the dashed lines in panels (b) and (c) labeled "Total". What the plots then clarify is the decomposition of the competing

[33] The required semiconductor material properties needed to compute mobilities are provided in Table A.2 in the Appendix.

scattering mechanisms at various doping densities, or temperatures.

Scattering rates: The scattering rates of all semiconductors are dominated by impurity scattering for the lowest energy carriers, and by a phonon branch at the highest energies. For compound semiconductors, the polar optical phonon (POP) is dominant at high energies, and for elemental semiconductors it is the acoustic deformation potential (or the optical deformation potential). What is striking is that at the lowest energies, the scattering rates dominated by impurity scattering are really not a whole lot different for various semiconductors: the mobility is therefore a very strong function of the carrier effective mass of the semiconductor. There are exceptions: for example, β-Ga_2O_3 has a rather large POP absorption rate, which leads to a relatively low mobility.

300 K mobilities vs. doping density: Panel (b) indicates that at the lowest doping densities, the intrinsic phonon-limited mobilities are attained, and at the highest doping densities Coulomb impurity scattering is dominant. Note that the conductivity $\sigma = qn\mu$, being a product, is still the highest for the highest doping because the decrease in the mobility is not as severe as the increase in the free carrier concentration. A larger dielectric constant helps screen Coulomb impurities more effectively as well, and since lower effective masses are typically linked with smaller bandgaps and larger dielectric constants, impurity scattering limited mobilities are high in InAs, Ge, and GaAs. On the other hand the low dielectric constant of diamond leads to a severe hit in the mobility at high doping. Note that the mobility increase starts saturating upon lowering the dopant concentration roughly an order below a critical density $N_{d,cr}$ given by $\mu_{imp} \sim \mu_{phonon}$ for the respective semiconductor, pointed by an arrow in panel (b) for InAs.

Mobilities vs. temperature for various doping densities: Panel (c) shows the trends of total mobilities vs. temperature. For low doping densities, the intrinsic phonon limited mobilities are observed near room temperature: therefore if there is a large increase in the mobility in a semiconductor upon lowering the temperature, it is a sign that it is not doped heavily. For such low-doped samples, the mobility is lowered significantly at low temperatures, if the carrier density is kept the same. Typically the density also changes with temperature: the carrier density freezes out depending on the activation energy of the dopants. On the other hand, for heavy doping the mobilities as well as the free carrier concentrations become nearly independent of temperature. This is a signature of the Mott transition, when the carrier density is degenerate with the Fermi level inside the band: and the behavior is nearly metallic. The highest room temperature mobilities are achieved in narrow-gap semiconductors: for the set shown, InAs achieves nearly $30,000$–$40,000$ $cm^2/V{\cdot}s$ at 300 K.

2D carrier mobilities: Fig. 23.40 shows the electron mobilities in AlGaAs/GaAs and AlGaN/GaN 2DEGs. Note that near-intrinsic polar optical phonon-limited mobilities are achieved at room temperature, even with high 2DEG densities because Coulomb impurity scattering

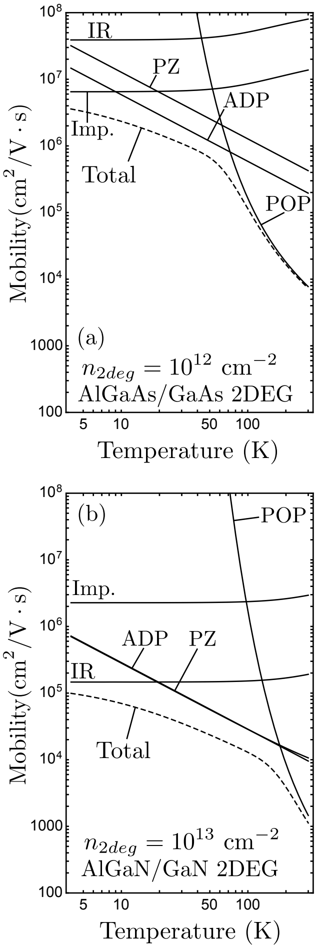


Fig. 23.40 Mobilities in (a) AlGaAs/GaAs and (b) AlGaN/GaN 2DEGs as a function of temperature. Note that for the most part they are similar to the phonon limited version that was shown in Fig. 23.38, except at the lowest temperatures. For the AlGaAs/GaAs 2DEG, an interface roughness characterized by $\delta = \frac{a_0}{2}$ and a lateral correlation length of $\Lambda = 50a_0$ is used, where a_0 is the lattice constant of GaAs. A delta-doping sheet of density $n_{imp} = 0.15 \times 10^{12}$ $/cm^2$ is placed 20 nm from the AlGaAs/GaAs 2DEG interface. For the AlGaN/GaN 2DEG, the interface roughness values used are $\delta = \frac{c_0}{2}$ and a lateral correlation length of $\Lambda = 20a_0$ are used, where (c_0, a_0) are the lattice constants of the wurtzite GaN lattice.

is heavily suppressed by spatially moving the scatterer dopants out of the way of the mobile electrons. This is not possible in the bulk because to obtain high electron densities high dopant concentrations are needed[34], which results in degradation of mobility by ionized impurity scattering as Fig. 23.39 (b) showed.

At the lowest temperatures, the highest achievable 2DEG mobilities are limited by either remote impurity scattering, or by interface roughness scattering. For example, the highest electron mobilities achieved in AlGaAs/GaAs 2DEGs at ~ 0.3 K today approach $\sim 44 \times 10^6$ cm^2/V·s, which requires scrupulous cleanliness, careful placement of remote dopants, and a very smooth interface. To highlight the effect of impurities and disorder, the plot shows a calculation of a case where these scattering mechanisms limit the mobilities to $\sim 4-5 \times 10^6$ cm^2/V·s, which, though not uncommon, is nearly 10 times lower than the records to date. Similarly, the experimental data for AlGaN/GaN 2DEGs are similar to what is shown in Fig. 23.40 (b). At the typically higher polarization-induced 2DEG densities in such structures, the interface roughness scattering mechanism dominates at low temperatures[35].

Electron mobilities in zero-gap 2D crystals such as graphene can be high because of the near-zero effective mass, but it is coupled with a vanishing energy gap. For 2D crystals with substantial bandgaps (e. g. transition metal dichalcogenides), typical room temperature mobilities are in the $100 - 1000$ cm^2/V·s range, and the highest low temperature mobilities are steadily improving, but are still far below 2DEGs in III-V semiconductors. With better purity of 2D crystals, the transport properties are expected to improve. Because of their unique geometry, the closeness of the surrounding dielectric layers to the mobile carriers in the 2D plane implies a rather large effect on the transport properties via large modifications of Coulomb and phonon scattering.

23.11 High-field velocity saturation

Fig. 23.41 (a) shows the trend of how the typical mobile carrier velocity changes with electric field, and Fig. 23.41 (b) shows the resulting characteristics measured experimentally in transistors. The velocity is linearly proportional to the electric field in the low field mobility regime. Beyond a characteristic field F_{sat}, the velocity saturates to $v_{sat} \sim 10^7$ cm/s in most semiconductors. If the field is further ramped up, free carriers cross the threshold of impact ionization or interband Zener tunneling, which will result in breakdown at characteristic fields F_{br}. This high-energy limit of tunneling and avalanche processes is the topic of discussion of Chapter 24. This section is dedicated to the discussion of the velocity in the field regime between $F_{sat} < F < F_{br}$ where the carrier velocity is saturated. This regime is of high importance in field-effect transistors, because as Fig. 23.41 (b) shows, in this regime the transistor output current becomes independent of the drain

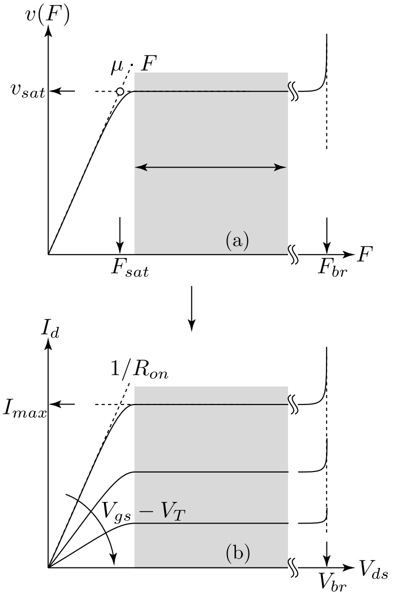


Fig. 23.41 Mobility, saturation velocity, and breakdown in semiconductors.

---

[34] An exception to this rule is the use of polarization-induced doping in graded heterostructures in semiconductors that have spontaneous and piezoelectric polarization, discussed in Chapter 15, Section 15.7.

[35] The low temperature transport properties are fundamentally important for such phenomena as the quantum Hall effect, obtaining resistance standards with parts per billion accuracy, and quantum oscillations of magnetotransport, phenomena that is discussed in Chapter 25. In applications of radio astronomy, cryogenically cooled HEMTs are used for imaging of distant galaxies because of their ultra-low noise stemming from the high electron mobilities. Furthermore, quantum computers based on microwave photons and superconducting Josephson junctions today require low-temperature low-noise amplifiers that can operate at the single photon level: these amplifiers are HEMTs based on III-V semiconductor heterostructures with 2DEG channels that have very high mobilities at low temperatures.

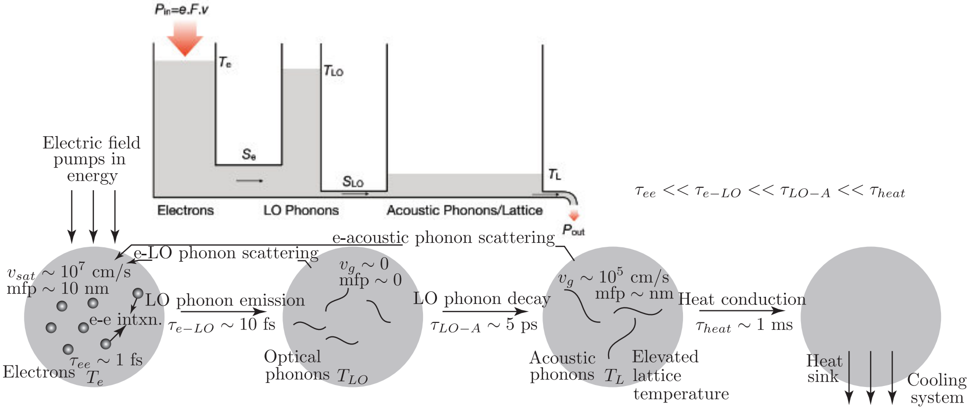

$P_{in}=e.F.v$
T_e
T_{LO}
S_e
S_{LO}
T_L
Electrons
LO Phonons
Acoustic Phonons/Lattice
P_{out}
$\tau_{ee} << \tau_{e-LO} << \tau_{LO-A} << \tau_{heat}$
Electric field pumps in energy
e-acoustic phonon scattering
e-LO phonon scattering
$v_{sat} \sim 10^7$ cm/s
mfp $\sim$ 10 nm
LO phonon emission
e-e intxn.
$\tau_{ee} \sim 1$ fs
Electrons T_e
$\tau_{e-LO} \sim 10$ fs
$v_g \sim 0$
mfp ~ 0
Optical phonons T_{LO}
LO phonon decay
$\tau_{LO-A} \sim 5$ ps
$v_g \sim 10^5$ cm/s
mfp $\sim$ nm
Acoustic phonons T_L
Elevated lattice temperature
Heat conduction
$\tau_{heat} \sim 1$ ms
Heat sink
Cooling system

Fig. 23.42 Energy flow and timescales in semiconductors with fast optical phonon scattering rates.

voltage, delivering gain and amplification.

Velocity saturation is a characteristic property of carriers at high electric fields in semiconductor crystals. Its origin is in the emission of phonons by high energy carriers at a fast rate. The scattering rates due to impurities and phonons combined with the semicondutor bandstructure allow the exact evaluation of velocity-field curves by numerical techniques[36].

To capture the essential physics of velocity saturation with a simple analytical model, we limit the discussion here by neglecting non-parabolicity and higher valleys. Invoking the laws of conservation of momentum and energy directly yield the velocity-field curves; the final results are rather simple to state. For a single-particle, the saturation velocity v_{sat} and saturation field F_{sat} are obtained as

$$\text{mom. cons.: } \frac{d(m^\star v)}{dt} = eF - \frac{m^\star v}{\tau_m} \text{ and energy cons.: } \frac{dE_{el}}{dt} = eFv - \frac{\hbar\omega_0}{\tau_E}$$

$$\frac{d(...)}{dt} = 0 \implies v_{sat} \sim \sqrt{\frac{\hbar\omega_0}{m^\star} \cdot \frac{\tau_m}{\tau_E}}, \text{ and } F_{sat} \sim \sqrt{\frac{\hbar\omega_0}{m^\star} \cdot \frac{(m^\star)^2}{e^2 \tau_m \tau_E}}, \tag{23.93}$$

where $\hbar\omega_0$ is the optical phonon energy, $m^\star$ is the carrier effective mass, and τ_m and τ_E are respectively the momentum and energy relaxation times. Momentum is gained from the field and lost in collisions. Similarly, energy is gained from the field and lost to the emission of phonons. Panel (a) of Fig. 23.39 shows that at high electron energies the momentum relaxation rate is a combination of acoustic phonons and ODP modes for elemental semiconductors (Ge, Si, diamond), but the emission rate into optical modes far dominates over all other modes for all compound semiconductors. The energy relaxation rates follow a similar trend, except the optical modes dominate because each emission results in far higher quanta of energy loss. Thus we can neglect the acoustic mode loss at high electric fields[37].

---

[36] The stochastic Monte Carlo method is especially well suited for this purpose since it allows a simple implementation of all scattering rates and the exact bandstructure that captures non-parabolicities and higher valleys. The further reading section refers to monographs dedicated to this topic. If high energy electrons climb to high-effective mass regions of the bandstructure due to band non-parabolicity, or to high energy high effective mass valleys, the velocity may actually drop with increasing field rather than stay constant. This phenomena and the resulting negative differential resistance was described for the Gunn effect in Chapter 17, Section 17.6.

[37] Equations 23.93 are only a part of the story. Fig. 23.42 indicates the energy flow *after* the emission into optical modes with analogy to water flow. The optical phonons decay into acoustic modes, which carry the energy away at the sound velocity to the heat sink. Each stage is characterized by its own timescale and effective temperature, causing both "hot"-electrons with $T_e >> T_L$ and "hot" optical phonons with $T_{LO} >> T_L$, where T_L is the physical lattice temperature, equal to the temperature of acoustic phonons.

For the crude approximation $\tau_E \sim \tau_m$, the expressions are $v_{sat} \sim \sqrt{\hbar\omega_0/m^\star} = v_0$ and $F_{sat} \sim \frac{v_0}{\mu}$, both of which are related to the optical phonon limited velocity $v_0 = \sqrt{\hbar\omega_0/m^\star}$ which is a characteristic property of a semiconductor, of the order $v_0 \sim 10^7$ cm/s. Because there are far more momentum scattering events than energy-changing (inelastic) events for carriers, $\tau_m < \tau_E$, the actual saturation velocities are typically lower than v_0. These considerations show that the highest saturation velocities are obtained for semiconductors with the smallest effective masses, and the largest optical phonon energies. We can immediately recognize that are competing requirements: small effective masses are typically connected to small energy bandgaps which are found in semiconductors that have heavy elements from the lower part of the periodic table such as InSb or InAs. Because of the small effective masses, the mobilities are high as well, but the semiconductors typically have small breakdown fields because of the small energy gaps. Furthermore, because of the heavy atoms, the optical phonon energy is rather small (~ 20 meV), and so is the thermal conductivity, and the ability to dissipate large amounts of heat[38].

From the above considerations, one may naively conclude that the saturation velocity is a fundamental property of the semiconductor. Both quantum mechanical considerations of Pauli blocking, and finite decay rate of phonons can significantly alter this conclusion, especially for low-dimensional carriers with high levels of carrier degeneracy, conditions that are routinely achieved today in semiconductor devices[39]. In the situations when high optical phonon emission rates are coupled with Pauli blocking, simplified analytical expressions are obtained for saturated current densities. A simple model may be developed by drawing analogy to ballistic current flow conditions. Assuming k_R is the Fermi wavevector of right-going carriers and k_L the Fermi wavevector of left-going carriers, the difference $k_R^2 - k_L^2 = k_{LO}^2$ where $\frac{\hbar^2 k_{LO}^2}{2m^\star} = \hbar\omega_{LO}$ is the threshold for phonon emission. Furthermore, for carriers in d-dimensions with equilibrium Fermi wavevector k_F, carrier number conservation requires that for parabolic bandstructure $k_R^d + k_L^d = 2k_F^d$ under conditions of high degeneracy. Using the expression for the current, Exercise 23.9 asks you to show that the saturation current density is given by

$$J_{sat} \approx \frac{e g_s g_v}{(4\pi)^{\frac{d+1}{2}}\Gamma(\frac{d+3}{2})}\frac{\hbar}{m^\star}(k_R^{d+1} - k_L^{d+1}) \implies$$

$$\boxed{J_{sat}^{1d} \approx e g_s g_v \frac{\omega_{LO}}{2\pi}}, \quad \boxed{J_{sat}^{2d} \approx e g_s g_v \frac{k_F}{\pi}\frac{\omega_{LO}}{2\pi}}, \text{ and } \boxed{J_{sat}^{3d} \approx e g_s g_v \frac{k_F^2}{8\pi}\frac{\omega_{LO}}{2\pi}}. \tag{23.94}$$

Under these conditions a) the effective saturation velocities can depend on the carrier concentrations, b) it is the *current density* that is saturated instead of the velocity, and c) a semiconductor with higher optical phonon energy will have a larger saturation current (Fig. 23.43).

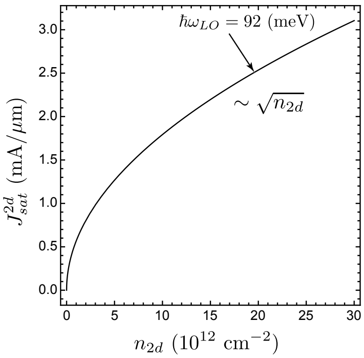


Fig. 23.43 Optical phonon emission limited saturation current model for 2DEGs. In this model the saturation velocity changes with the mobile carrier density.

---

[38] On the other hand, for semiconductors with light atoms from the top of the periodic table (e.g. GaN, SiC, diamond, AlN), the optical phonon energies are large (~ 100 meV), but the bandgaps and the effective masses are also large. These considerations end up giving the narrower bandgap semiconductors only a slight advantage of higher saturation velocities, but at the cost of a significantly reduced breakdown field and low thermal conductivity. For example, InAs has $v_{sat} \sim 3 \times 10^7$ cm/s, $F_{sat} \sim 1$ kV/cm, $F_{br} \sim 40$ kV/cm, and thermal conductivity $\kappa \sim 27$ W/m·K compared to GaN which has $v_{sat} \sim 10^7$ cm/s, $F_{sat} \sim 100$ kV/cm, $F_{br} \sim 3500$ kV/cm, and $\kappa \sim 280$ W/m·K. These considerations dictate the different regimes of usage for devices: for ultrafast transistors InAs is preferable, but is limited in the voltage and power it can deliver. GaN delivers more power at lower speeds.

[39] As an example, the saturation *current* in 1D carbon nanotubes is observed to be nearly $I_{sat} \sim 4 \cdot e \cdot \frac{\omega_{LO}}{2\pi}$, *independent* of the carrier density in the high-density limit. Therefore, the effective saturation velocity for a 1D carrier density n_{1d} is $v_{sat} = I_{sat}/(en_{1d})$, inversely proportional to the carrier density.

23.12 Chapter summary section

In this rather long chapter, using quantum mechanical transition rates we explicitly derived the carrier mobilities in various semiconductors. We learned:

- Typical trends of carrier mobilities in semiconductor bulk and lower-dimensional structures,
- The method to evaluate carrier transport properties from microscopic scattering mechanisms,
- How to account for the scattering by multiple scatterers, and for distributions of carriers, and for screening of scattering potentials,
- Scattering rates and resulting carrier mobilities due to point defects, Coulomb impurities, dipoles, dislocations, alloy disorder, and interface roughness for both 3D and 2D carriers,
- Scattering rates and resulting carrier mobilities for phonons for four modes: acoustic deformation potential (ADP), acoustic piezoelectric (PZ), optical deformation potential (ODP), and polar optical phonons (POP), for both 3D and 2D carriers,
- The net effect of the various scatterers in several semiconductor families, which highlighted that in pure semiconductors the high-temperature mobilities are limited by phonon scattering and by impurities and disorder at low temperatures, and finally
- That at high electric fields the velocity deviates from a mobility-limited, linear regime and the velocity saturates due to optical phonon emission. The strong effect of optical phonons on the saturation velocity of semiconductors were discussed.

Further reading

The problem of carrier transport in both metals and semiconductors is treated rigorously in **Carrier Scattering in Metals and Semiconductors** by Gantmakher and Levinson with a wealth of insights, but it is limited to 3D bulk structures. **Fundamentals of Carrier Transport** by M. Lundstrom provides a more modern treatment for semiconductors in 3D and 2D, and describes several levels of approximations. **Semiconductor Physics: An Introduction** by K. Seeger and **Basic Semiconductor Physics** by C. Hamaguchi have excellent treatments of transport as part of a comprehensive exposition of the physics of semiconductors. **Wave Mechanics applied to Semiconductor Heterostructures** by G. Bastard and **Quantum Wells, Wires and Dots** by P. Harrison have detailed treatments of transport in semiconductor heterostructures in low dimensions. **Electrons and Phonons** by J. Ziman is an early and classic text, and **Electrons and Phonons in Semiconductors Multilayers** by B. Ridley and **Phonons in Nanostructures** by M. Stroscio and M. Dutta are more modern treatments for heterostructures. Finally, **The Monte Carlo Method for Semiconductor Device Simulation** by C. Jacoboni and P. Lugli is an excellent treatment of this technique applied to semiconductors for exploring high-field transport phenomena.

Exercises

(23.1) **Electrical resistivity of metals**

Consider a heavily degenerate n-type carrier concentration of carrier concentration n and Fermi wavevector k_F doped with n_{imp} ionized impurities.

(a) Calculate and show that if the impurity potential is strongly screened, the electrical resistivity is $\rho = \frac{m}{nq^2\tau} = \frac{\pi^2}{2} \frac{h}{q^2} \frac{1}{k_F} \frac{n_{imp}}{n}$.

(b) Using the model of heavily screened ionized impurity scattering, show that the electrical resistivity of a metal should increase at the rate of $\Delta\rho \approx 1\mu\Omega \cdot \text{cm}$ for every 1% increase in the impurity density.

(23.2) **Spin and valley degeneracies in scattering**

In the earlier chapters when we calculated electrical currents, the spin and valley degeneracies always appeared in the sums. But in the application of the Fermi's golden rule, we do not include these degeneracies in the calculations. This is due to physical consideration of the scattering processes, which are discussed in this problem.

(a) None of the scattering mechanisms discussed in this chapter can flip the electron spin in an ordinary manner. Why is this so? What sort of scatterers can flip the spin of the electrons? Argue why this implies that spin-up electrons scatter into spin-up states, and spin-down into spin-down states, and they don't mix. This is why the spin-degeneracy is not included.

(b) Spin-orbit interaction in semiconductors opens up the channels of spin-scattering even for ordinary scatterers of part (s). This is manifested in the Dyakonov–Perel, the Elliott–Yafet, and the Dresselhaus scattering mechanisms. Read about these mechanisms and comment how the rates of such scattering compare to the ones discussed in this chapter.

(c) Discuss why intervalley scattering by phonon absorption is typically forbidden when only phonons of low momentum are available. But argue why intervalley scattering is allowed for electrons with high-enough energies. Discuss the required criteria, and why semimetals specifically have significant scattering of this type.

(23.3) **Screening and the Lindhard function**

(a) Using the Lindhard screening function, prove the results that appear in Table 23.1. Discuss screening in 1D, which is not covered in this chapter explicitly.

(b) Much of the original screening models were developed for ions in solutions in electrochemistry. Study the analogy of screening by ions in liquid solutions to that by mobile electrons or holes in semiconductors and discuss the similarities and differences.

(23.4) **The algebra of the Fang–Howard–Stern variational wavefunctions**

(a) In Table 23.2 we listed the various form factors for scatting in 2D electron systems of different kinds. Device the results of the table. Then, show that the matrix elements for both the Fang–Howard–Stern type variational wavefunction and the ∞-Well reduce to that of the δ-function well in the limit of strong quantum confinement.

(b) Prove the results of alloy scattering given in Equation 23.62.

(c) Show from Equations 23.62 that the centroid of the modified Fang–Howard–Stern function in the well side is given by $\langle z \rangle = \frac{6N^2}{b^4}(1 + \frac{2}{3}bz_0 + \frac{1}{6}b^2z_0^2 - \frac{z_0^2 b^4}{6\kappa_b^2})$.

(23.5) **Similarity of the formulations of point-defect and alloy scattering**

The results of point-defect scattering in Section 23.3 and of alloy disorder scattering in Section 23.7 have several similarities. Closer inspection would show that they are identical: discuss this correspondence, and show that all results of alloy scattering could have been obtained directly from the results of point-defect scattering.

(23.6) **Interface roughness scattering: integrals and choice of approach**

In Section 23.8 we discussed scattering of electrons due to roughness of interfaces. Answer the following on this scattering mechanism:

(a) Prove that for the exponential correlation function $\delta^2 e^{-(\frac{\rho-\rho'}{\Lambda})}$, $\int_{\rho=0}^{\infty} d\rho \int_{\rho'=0}^{\infty} d\rho' \cdot \rho \cdot \rho' \cdot [\Delta(\rho) \cdot \Delta(\rho')] \cdot J_0(q_{\|}\rho) \cdot J_0(q_{\|}\rho') = \frac{\delta^2 \Lambda^4}{[1+(q_{\|}\Lambda)^2]^3}$.

(b) Prove that for the Gaussian correlation function $\Delta(\rho)\Delta(\rho') = \delta^2 e^{-(\frac{\rho-\rho'}{\Lambda})^2}$, $\int_{\rho=0}^{\infty} d\rho \int_{\rho'=0}^{\infty} d\rho' \cdot \rho \cdot \rho' \cdot [\Delta(\rho) \cdot \Delta(\rho')] \cdot J_0(q_{\|}\rho) \cdot J_0(q_{\|}\rho') = \frac{\delta^2 \Lambda^4 e^{-\frac{(q_{\|}\Lambda)^2}{2}}}{4}$.

(c) Discuss why in standard silicon MOSFETs, the effective field-approach to evaluating interface roughness scattering is the correct approach (Hint: there is no physical quantum well of thickness L_z).

(23.7) **Optical deformation potential emission**

Using the angular relations shown in Fig. 23.33 and following the same method as Equation 23.76 for the corresponding absorption process, show that the relaxation rate for optical deformation potential scattering is $\frac{1}{\tau_{m,em}^{ODP}(\mathbf{k})} = \frac{\pi D_{op}^2 (N_{LO}+1)}{2\rho\omega_{LO}} \cdot \frac{1}{2\pi^2} (\frac{2m^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{E_{\mathbf{k}} - \hbar\omega_{LO}} \cdot \Theta(E_{\mathbf{k}} - \hbar\omega_{LO})$ where $\Theta(...)$ is the unit step function.

(23.8) **Piezoelectric and Polar Optical Phonon scattering in 2D systems**

In Table 23.2, the last two rows present exact and approximate forms for $\mathcal{G}(\beta_{\|})$.

(a) Verify the integrals by direct evaluation and show they reduce to the approximations.

(b) Examine the accuracy of the approximate form, and show that it deviates less than a few percent over most practical range of wavevectors $0 \leq q_{\|} \leq 2k_F$.

(c) Use the approximate form to obtain the long-wavelength (low energy, $\beta_{\|} \to 0$) limit and the short-wavelength (high-energy) limits of this factor. What are the implications of these limits for polar scattering of low- and high-energy 2D electron or hole gases?

(23.9) **Phonon-limited saturation current densities**

(a) Prove the results stated in Equation 23.94 for the phonon-limited saturation current densities.

(b) Make plots of these current densities for several semiconductors such as GaAs, GaN, and β-Ga_2O_3.

(c) If the saturation current is given by Equation 23.94 for a semiconductor channel of a FET, find the dependence of the transconductance $g_m^{sat} \sim \frac{\partial J_{sat}}{\partial V_{gs}}$ of the corresponding FET, and discuss how it differs from the case when the current is limited by a constant saturation velocity $J_{sat} \sim qnv_{sat}$.

(d) Explore if this simplified model indeed explains experimental data by comparing with reported saturation currents.

(23.10) **Electron mobility in ultrathin quantum wells**

Consider a 2D electron gas located in the ground state of the conduction band of a heterostructure quantum well of area A as sketched in Fig. 23.44. Assume for this problem that the band offset is infinite. The electron is free to move in the $x-y$ plane, and its motion is quantized in the z-direction.

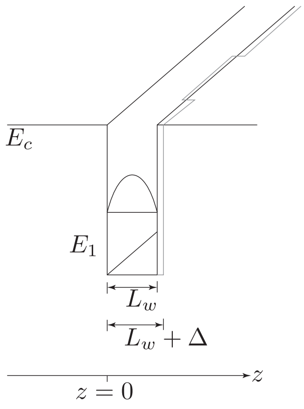


Fig. 23.44 A quantum well with rough interfaces.

(a) Write the total wavefunction in the effective mass approximation of the electron in the lowest (or ground-state) subband as a function of the lateral wavevector $\mathbf{k} = (k_x, k_y)$, the lateral real-space coordinate $\mathbf{r} = (x, y)$ and area A, the quantum well width L_w and the z-coordinate, and the quantum well lattice-periodic function $u_c(r)$. Identify the envelope function and retain it for the rest of the sections of this problem.

(b) Write the effective mass equation for the quantum well of width L_w and show that the energy eigenvalues are given by $E = \frac{\hbar^2}{2m_c^\star}(k_x^2 + k_y^2) + E_1(L_w)$ where $E_1(L_w)$ is the ground state energy of the quantum well.

(c) Show that the ground state energy of the quantum well depends on the quantum well width as $E_1(L_w) = \frac{h^2}{8m_c^\star L_w^2}$.

(d) Due to fluctuations in the epitaxial growth of the quantum well, the well width was not exactly L_w everywhere in the (x, y) plane, but fluctuated by a thickness $\Delta(x, y)$ such that the fluctuation is much smaller than the well width $\Delta(x, y) << L_w$ and the net fluctuation is zero $\langle \Delta(x, y) \rangle = 0$ as shown in gray lines in Fig. 23.44. Show then that the ground state energy fluctuates by $\Delta E_1(x, y) = -\frac{h^2}{4m_c^\star L_w^3}\Delta(x, y)$.

(e) Show using the effective mass equation that this fluctuation in the ground state energy due to well-width variations can be treated as a perturbation $\Delta E_1(x, y) = W(x, y) = W(\mathbf{r})$ in the motion of the electron in the $x - y$ plane.

(f) The scattering rate for the 2DEG electrons $\frac{1}{\tau(\mathbf{k} \to \mathbf{k}')} = \frac{2\pi}{\hbar}|\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle|^2 \delta(E_\mathbf{k} - E_{\mathbf{k}'})$ due to these fluctuations from Fermi's golden rule depends on the square of the matrix element $\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle$. Show that $|\langle \mathbf{k}'|W(\mathbf{r})|\mathbf{k}\rangle|^2 = \frac{h^4}{16(m_c^\star)^2 L_w^6} \cdot |\int \frac{d^2r}{A} e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{r}}\Delta(x, y)|^2$. There is no need to evaluate the integral.

(g) Because the electron mobility is $\mu = \frac{q\langle\tau\rangle}{m_c^\star}$, show that the mobility limited by this quantum well interface roughness (IR) scattering decreases as the *sixth power* of the quantum well width according to $\mu_{IR} \propto \frac{q m_c^\star L_w^6}{h^3}$. Show this proportionality, there is no need to evaluate integrals. This is a severe scattering mechanism that is a hurdle to achieving high electron mobilities for 2DEGs in very thin quantum wells, because $\mu_{IR} \downarrow$ as $L_w \downarrow$ as the sixth power.

Through the Barrier: Tunneling and Avalanches

24

In this chapter we explore the fascinating phenomena of quantum tunneling of electrons in semiconductors and their quantized heterostructures. Our specific goals are to understand:

- In what ways are tunneling phenomena in semiconductors similar to, or different from tunneling in atomic physics where it was first discovered?
- How can we quantitatively calculate currents in semiconductors due to intraband and interband tunneling?
- How is tunneling modified in lower dimensions?
- What are the applications of tunneling in semiconductor electronics and photonics?
- How do avalanche processes and interband tunneling cause electrical breakdown in semiconductors?

24.1 Tunneling: a résumé 611
24.2 Single-barrier tunneling 613
24.3 WKB tunneling theory 615
24.4 WKB for semiconductors 617
24.5 Nordheim supply function 621
24.6 Fowler–Nordheim tunneling 621
24.7 Interband Zener tunneling 624
24.8 pn tunnel junctions in 1D, 2D, and 3D 624
24.9 NDR, backward diodes 629
24.10 Tunneling FETs 633
24.11 Resonant tunneling 634
24.12 Bardeen's tunneling theory 635
24.13 Kubo formalism 636
24.14 Landau–Zener theory 637
24.15 Avalanche processes and impact ionization 639
24.16 Tail of the dragon 642
24.17 Chapter summary section 642
Further reading 643
Exercises 643

24.1 Tunneling: a résumé

The history of quantum tunneling goes back to times well before the electron was discovered or quantum mechanics was developed. Experimentally, the discovery of **radioactivity** by Becquerel, Marie Curie (Fig. 24.1), and Pierre Curie was among the first observations of the effect of quantum tunneling. Rutherford's careful investigation of radioactivity led to the identification of the mechanism of **alpha–decay**, in which the heavy nucleus spits out alpha particles (He^{2+} nuclei). Though experimentally identified in 1907, its explanation had to wait for nearly two decades for the development of quantum mechanics.

In this intervening time, strange phenomena related to the emission of electrons from metals into vacuum were experimentally observed. Richardson and other researchers had already observed the phenomena of thermionic emission, as was discussed in Chapter 18: the electron current that could be extracted from the metal increased with temperature as $J = A^\star e^{-q\phi_b/k_bT}(e^{qV/k_bT} - 1)$. But Lilienfeld and other experimentalists observed another component that had a much weaker dependence on temperature, called **cold field–emission**. The emission of electrons was enhanced at large electric fields, even when the metal was kept cold. This phenomenon had no classical explanation.

Fig. 24.1 Marie Curie discovered radioactivity with Pierre Curie and Becquerel, and also discovered the elements radium and polonium. She was awarded two Nobel Prizes, one in Physics and the other in Chemistry. Quantum tunneling enables radioactive decay processes.

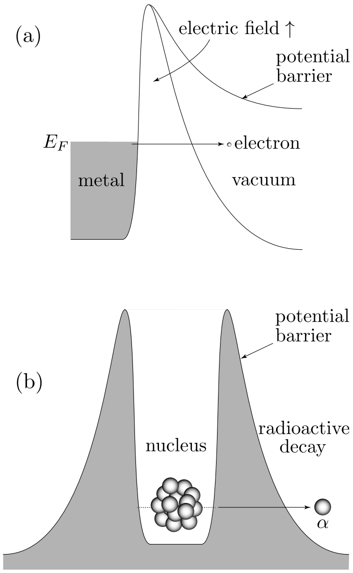


Fig. 24.2 (a) Field–emission from metals, and (b) Radioactive decay.

Fig. 24.3 George Gamow offered the first explanation of α-decay as the process of quantum tunneling of Helium nuclei out of the larger radioactive nuclei. He wrote several popular physics books and famously an important cosmology paper with the author list Alpher, Bethe, Gamow' as a practical joke.

Right after Pauli identified the exclusion principle, and Fermi and Dirac developed the resulting distribution function for fermions (Chapters 3 and 4), Sommerfeld proposed that the Fermi–Dirac statistics is the correct one for electrons in metals (Chapter 5). Quantum mechanics was developed in the form of Schrödinger's wave equation around 1927. This immediately led to the resolution of the earlier mysterious phenomena related to tunneling.

The explanation of cold field emission by Ralph Fowler and Lothar Nordheim used the combination of Fermi–Dirac statistics of electrons in metals, with the tunneling of electron waves from the metal into vacuum when the potential barrier is made thinner (*not* smaller) with an electric field. This is shown schematically in Fig. 24.2 (a). This explanation offered strong support to Sommerfeld's assertion of the application of Fermi–Dirac statistics to electrons in metals. It simultaneously gave a vivid example of the wave nature of electrons and confirmation of the validity of the electron wavefunction in Schrödinger's equation. Robert Oppenheimer around the same time evaluated the rate of ionization of atoms in the presence of electric fields using the Schrödinger equation (Fig. 24.4), analogous to the Fowler–Nordheim theory for metals. We will study **Fowler–Nordheim tunneling** in several semiconductor structures in this chapter.

The explanation of α-decay in radioactivity by Gamow (Fig. 24.3) and by Gurney and Condon followed essentially a very similar model, but now for the tunneling of blocks of neutrons and protons as helium nuclei from the bigger radioactive nucleus, as indicated in Fig. 24.2 (b). Lev Landau and Clarence Zener independently developed the Landau–Zener model for tunneling in adiabatic processes in the early 1930s. Specifically, in 1934 Zener developed a model for the electrical breakdown of insulators and dielectrics at high electric fields by proposing the quantum tunneling of electrons from the valence band into the conduction band across the energy bandgap. The idea of energy bands and gaps were very new at that time. **Zener tunneling** will be discussed later in this chapter.

In 1934, using microwave spectroscopy it was discovered that the nitrogen atom in an ammonia molecule (NH_3) tunnels from one side of the triangle formed by the hydrogen atoms to the other. The frequency of this tunneling ($\sim$ 24 GHz) is in the microwave regime, and was crucial for the development of the microwave amplifier by stimulated emission of radiation, or the ammonia **MASER**, the precursor of the laser. The precision of this tunneling frequency made it the first **atomic clock** for precision timekeeping guaranteed by the laws of quantum mechanics.

Electrical breakdown effects were observed in semiconductors in the 1950s. The most striking effect of tunneling was observed by Leo Esaki in 1958 in heavily doped germanium pn junction diodes. This first observation of controlled quantum tunneling in an all- semiconductor junction produced a negative differential resistance (NDR). The NDR behavior (see Fig. 24.5) allowed for sustained high-frequency oscillations and generation of microwave power from such **Esaki diodes**, which will be discussed later in the chapter. Esaki's observation of electron tunneling in a semiconductor junction was the first in a condensed matter system, and was the precursor to the observation of analogous behavior in superconductors by Giaever. The tunneling of Cooper pairs was predicted by Brian Josephson in 1962, and observed by Fiske, Rowell, and Anderson within a year. Such **Josephson junctions** are used for microwave qubits for quantum computation.

Esaki and Tsu proposed the creation of artificial semiconductor superlattices to controllably design the Gunn diode behavior (Chapter 17, Section 17.6). These ideas led to the realization of double-barrier **resonant-tunneling diodes** (RTDs) which exhibit strong NDR in both positive *and negative* bias conditions, unlike the pn tunnel junction which only shows such behavior in the forward-bias. Tunneling transport in semiconductor RTDs is analogous to the **Ramsauer–Townsend effect** in atomic physics, in which electrons of certain energies can transmit through molecules of a gas with near unity transmission coefficient without scattering, as if the atoms did not exist.

Tunneling phenomena in semiconductors play a major role in today's science and technology. **Scanning tunneling microscopy** (STM) is routinely used to study the surface electronic states of semiconductors, and **transmission electron microscopy** (TEM) and **scanning electron microscopes** (SEMs) are routinely used to characterize and image semiconductor materials and devices. STM, SEM, and TEM all use the tunneling of electrons as the basis of their operation. **High-frequency oscillators** are based both on interband and intraband resonant tunneling. Low-resistance **ohmic contacts** for energy-efficient and high-speed transistors employ tunnel junctions of metals with heavily doped semiconductors. Low-resistance interband tunneling junctions are used in **multijunction solar cells** to connect pn-junctions of different energy bandgaps in series (Chapter 28). Most high-efficiency **lasers** and **light-emitting diodes** take advantage of tunnel junctions for controlling the flow of current and for efficient current spreading (Chapter 29). The process of tunneling followed by photon emission is the mechanism of operation of **quantum cascade lasers**. Tunneling of electrons across very thin insulating gate barriers in MOSFETs (Chapter 19) posed a problem in scaling, forcing a move towards high-K dielectrics and thicker barriers. Making the subthreshold slope of transistors steeper than the standard 60 mV/decade increases their energy efficiency. Interband tunneling is being investigated as a potential route to achieve this in **tunneling field-effect transistors** (TFETs).

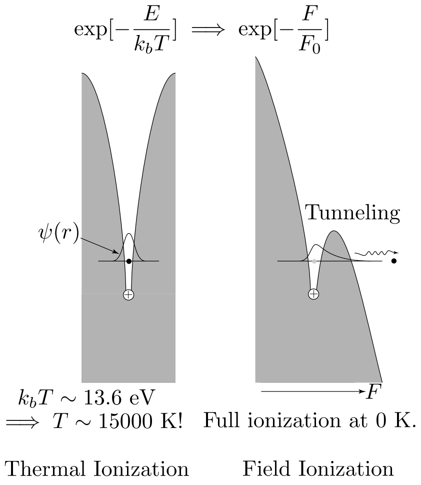


Fig. 24.4 Field–ionization of atoms by tunneling. While enormous heat energy is needed to excite an electron from the ground state of hydrogen to the continuum, a relatively modest electric field bends the potential barrier and allows the electron to tunnel out.

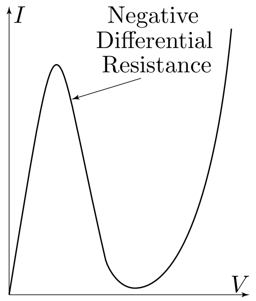


Fig. 24.5 Characteristic shape of negative differential resistance (NDR) in current-voltage characteristics observed in Esaki (pn) diodes and in resonant tunneling diodes.

24.2 Single-barrier tunneling

Fig. 24.6 shows the transmission of an electron over a barrier, and the quantum mechanical tunneling of an electron through the barrier. For a free–electron of mass m_e and kinetic energy E, the boundary

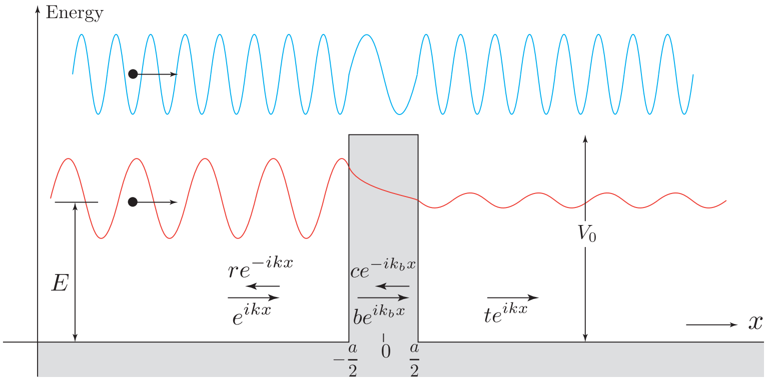


Fig. 24.6 Tunneling and transmission of electrons through a single potential barrier.

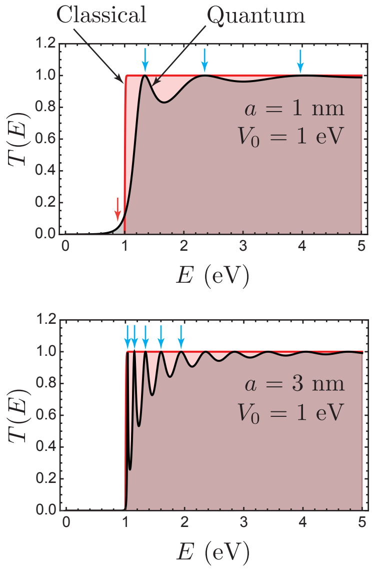


Fig. 24.7 Transmission probability $T = |r|^2$ as a function of the electron energy for a barrier height $V_0 = 1$ eV, and two thicknesses. For electron kinetic energies larger than the barrier height, the transmission probability is given by Equation 24.1, and the electron wavevector is real at all x as shown in the wave in blue in Fig. 24.6 and the transmission probability is close to one for high energies, with some resonant reflections. For electron kinetic energies lower than the barrier height, the tunneling transmission probability is given by Equation 24.2, and as shown in red in Fig. 24.6, the amplitude of the function decays exponentially in the barrier and a small amplitude leaks through to the other side.

conditions at the two barrier interfaces $x = \pm a/2$ give us four equations and four unknowns: the coefficients r, b, c, and t as indicated in the figure. Using boundary conditions that the wavefunction and its derivative are continuous at $x = +a/2$ and at $x = -a/2$ we get four equations. Solving these four equations yields the four unknown coefficients, solving the problem exactly.

Since we are interested in the tunneling probability $T(E) = |t|^2$, the above calculation yields

$$\boxed{T(E) = \frac{1}{1 + \frac{V_0^2}{4E(E-V_0)}\sin^2(k_b a)}.} \tag{24.1}$$

Here $k_b = \sqrt{\frac{2m_e}{\hbar^2}(E - V_0)}$ is the electron wavevector in the barrier region, with the meaning $E = \frac{\hbar^2 k_b^2}{2m_e}$ is the kinetic energy. The kinetic energy can be lower than the barrier height V_0, in which case we still can write the total energy as the sum of kinetic and potential energies, $E = \frac{\hbar^2 k_b^2}{2m_e} + V_0$. For this to be consistent, when the kinetic energy of the electron is lower than the barrier ($E < V_0$), k_b has to be imaginary, and $\sin(iy) = i\sinh(y)$. Fig. 24.7 shows plots of the tunneling probability as a function of the electron kinetic energy E for a fixed barrier height $V_0 = 1$ eV, and barrier thicknesses of $a = 3$ nm and $a = 1$ nm. Note how sensitive tunneling is to thickness, and the resonances when $E > V_0$. These values are routinely encountered in semiconductor devices and nanostructures. For $E < V_0$ the tunneling probability is

$$\boxed{T(E) = \frac{1}{1 + \frac{V_0^2}{4E(V_0-E)}\sinh^2(k_b a)} \approx \frac{16E(V_0 - E)}{V_0^2}e^{-2k_b a}.} \tag{24.2}$$

The boxed expression above is the *exact* tunneling probability for the rectangular barrier, and the approximation will be seen to be equivalent to the WKB approximation, which is the technique that can be applied to potential barriers or arbitrary shapes.

24.3 WKB tunneling theory

The WKB method solves the time-independent Schrödinger equation for arbitrary potentials. There is typically a misconception that this method is a crude approximation – far from it[1]! Consider the time-independent Schrödinger equation in 1D:

$$-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}\psi(x)+V(x)\psi(x)=E\psi(x), \tag{24.3}$$

rewritten in the form

$$\psi''(x)=Q(x)\psi(x), \tag{24.4}$$

with

$$Q(x)=\frac{2m_e}{\hbar^2}(V(x)-E). \tag{24.5}$$

This is in general analytically unsolvable, except for few well-known cases such as the particle in a box, the Delta function, the harmonic oscillator, or the hydrogen atom. George Green (the same person behind Green's functions) had a brilliant insight for an asymptotic solution. He suggested that we try a solution of the form

$$\psi(x)=e^{S(x)}. \tag{24.6}$$

Substituting, we obtain $(S''+(S')^2)e^S=Qe^S$ which leads to

$$S''+(S')^2=Q. \tag{24.7}$$

Note that this is *exact*. But unfortunately we have just converted the already unsolvable Schrödinger equation into another insolvable equation[2]. It is also a non-linear equation!

Green made the crucial observation that if $S(x)=ax^b$ where $b<0$, then $S''<<(S')^2$ for small x. This makes the asymptotic form of Equation 24.7:

$$(S')^2\sim Q, \tag{24.8}$$

which leads to

$$S'\sim\pm\sqrt{Q}, \tag{24.9}$$

and

$$S(x)\sim\pm\int_a^x du\sqrt{Q(u)}. \tag{24.10}$$

Here a is any chosen constant. Now lets write

$$S(x)=\underbrace{\pm\int_a^x du\sqrt{Q(u)}}_{S_0(x)}+\underbrace{C(x)}_{S_1(x)+S_2(x)+S_3(x)+...}, \tag{24.11}$$

where we have gone back from an asymptotic form to an exact form. Note that the leading functional form is defined $S_0(x)=\pm\int_a^x du\sqrt{Q(u)}$. Consequently, we have the relations

$$S_0'(x)=\pm\sqrt{Q(x)}, \tag{24.12}$$

---

[1] This section may be skipped at a first reading. The main WKB result is highlighted in Equation 24.25; it should be understood and carried forward. This section discusses the mathematical derivation of the WKB approximation of the Schrödinger equation. It closely follows the excellent lectures by Carl Bender available here: `https://tinyurl.com/54ckyef9`.

[2] This form of a differential equation is called the Riccati equation.

and

$$S_0''(x) = \pm\frac{Q'(x)}{2\sqrt{Q(x)}}. \tag{24.13}$$

Now we substitute $S = S_0 + C$ into 24.7 to get

$$S_0'' + C'' + (S_0' + C')^2 = Q. \tag{24.14}$$

Note that this is back to being *exact*. Writing it out, we get

$$S_0'' + C'' + (S_0')^2 + (C')^2 + 2S_0'C' = Q, \tag{24.15}$$

and substituting all terms related to $S_0(x)$ and its derivatives, we get

$$\pm\frac{Q'(x)}{2\sqrt{Q(x)}} + C'' + Q + (C')^2 \pm 2\sqrt{Q(x)}C' = Q. \tag{24.16}$$

This is still *exact*. Note the crucial cancellation of Q. The result is a differential equation, but *now for the correction function* $C(x)$:

$$\pm\frac{Q'(x)}{2\sqrt{Q(x)}} + C'' + (C')^2 \pm 2\sqrt{Q(x)}C' = 0. \tag{24.17}$$

But this equation is not very different from the Riccati equation either! Now we step back to asymptotics again: since $S_0(x)$ is the leading order or dominant function, we have the relations $C(x) << \pm\int_a^x dt\sqrt{Q(t)}$, and consequently

$$C'(x) << \pm\sqrt{Q(x)}, \tag{24.18}$$

and

$$C''(x) << \pm\frac{Q'(x)}{2\sqrt{Q(x)}}. \tag{24.19}$$

As a result, we may throw away the C'' and $(C')^2$ terms to obtain:

$$C'(x) \sim -\frac{Q'(x)}{4Q(x)}, \tag{24.20}$$

which yields

$$C(x) \sim -\frac{1}{4}\ln Q(x) = \underbrace{-\frac{1}{4}\ln Q(x)}_{S_1(x)} + S_2(x) + S_3(x) + \ldots. \tag{24.21}$$

Thus, the solution to the Schrödinger equation is

$$\psi(x) = e^{S(x)} = e^{S_0(x)+S_1(x)+S_2(x)\ldots} = e^{S_0(x)}e^{S_1(x)}e^{S_2(x)+\ldots}. \tag{24.22}$$

Substituting the values of S_0 and S_1, we obtain

$$\psi(x) = e^{S(x)} = e^{\pm\int_a^x du\sqrt{Q(u)}}e^{-\frac{1}{4}\ln Q(x)}e^{S_2(x)+\ldots}, \tag{24.23}$$

which is rewritten in the form

$$\psi(x) = e^{S(x)} = \frac{1}{(Q)^{\frac{1}{4}}} e^{\pm \int_a^x du \sqrt{Q(u)}} e^{S_2(x) + \ldots}. \tag{24.24}$$

This process may be repeated again and again. Each step will generate new functional forms $S_2(x), S_3(x)$.... The series becomes rational powers of x: $D(x) = S_2(x) + S_3(x) + \ldots = \sum_n a_n(\sqrt{x})^n$. In principle, the functions in the exponent $S_2(x), S_3(x)$... can recover the full functional form of $\psi(x)$.

For most cases, it turns out that $D(x) << 1$, and the "WKB" approximation to use looks like

$$\boxed{\psi(x) \approx \frac{K}{Q(x)^{\frac{1}{4}}} e^{\pm \int_a^x du \sqrt{Q(u)}},} \text{ where } \boxed{Q(x) = \frac{2m_e}{\hbar^2}(V(x) - E),} \text{ and} \tag{24.25}$$

where K is a constant.

One can *mechanize* the extraction of the successive functions $S_n(x)$ by the following *perturbative* approach. We recast the Schrödinger equation as a function of a parameter ϵ of the form $\epsilon^2\psi'' = Q\psi$, and assume a solution

$$\psi(x) = e^{S} = e^{\frac{1}{\epsilon}\sum_{n=0}^{\infty} S_n \epsilon^n} = e^{\frac{1}{\epsilon} S_0 + S_1 + S_2\epsilon + S_3\epsilon^3 + \ldots}. \tag{24.26}$$

Substituting this form into the Schrödinger equation $\epsilon^2\psi'' = Q\psi$ leads to the set of equations defined by

$$(S_0' + S_1'\epsilon + S_2'\epsilon^2 + S_3'\epsilon^3)^2 + (S_0''\epsilon + S_1''\epsilon^2 + S_2''\epsilon^3 + \ldots) = Q, \tag{24.27}$$

from where we equate like powers of ϵ on both sides of the equation to obtain

$$\begin{aligned}
(S_0')^2 &= Q \quad \text{for } \epsilon^0 \tag{24.28}\\
2S_0'S_1' + S_0'' &= 0 \quad \text{for } \epsilon^1 \tag{24.29}\\
2S_0'S_2' + S_1'' + (S_1')^2 &= 0 \quad \text{for } \epsilon^2 \tag{24.30}\\
2S_0'S_n' + S_{n-1}'' + \sum_{j=1}^{n-1} S_j'S_{n-j}' &= 0 \quad \text{for } \epsilon^n. \tag{24.31}
\end{aligned}$$

What is the small parameter ϵ? One way to look at it is to "relate" it to $\hbar$, which is a small number. But ϵ is a variable in the above formulation, which renders the Schrödinger equation into a polynomial series, enabling the WKB evaluation. We leave this section and carry forward Equation 24.25 as the main WKB result.

24.4 WKB for semiconductors

Now consider the effective mass equation for an electron in the conduction band of a semiconductor where the band edge $E_c(x)$ varies

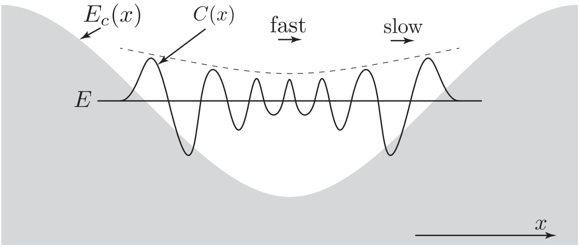


Fig. 24.8 Electron transport and approximate wavefunction in smoothly varying potentials.

smoothly in space. Then, for the band $E_n(k) = E_c(x) + \frac{\hbar^2 k_x^2}{2m_c^\star}$, the effective mass equation for the electron is

$$[E_n(-i\nabla)]C(x) = [E]C(x) \implies \frac{d^2}{dx^2}C(x) = \underbrace{-\frac{2m_c^\star}{\hbar^2}[E - E_c(x)]}_{Q(x)} C(x), \tag{24.32}$$

where $C(x)$ is the envelope function, and the full wavefunction of the wavepacket representing the electron is $\psi(x) \approx C(x)u_k(x)$. This is clearly in the form of the WKB formulation; the solution is

$$\boxed{C(x) \approx \frac{K}{Q(x)^{\frac{1}{4}}} e^{\pm \int_a^x du \sqrt{Q(u)}},} \tag{24.33}$$

where K is a constant. Now if the energy of the wavepacket is larger than the conduction band edge $E > E_c(x)$, $Q(x) < 0$, and $\sqrt{Q(x)} = ik(x)$, where $k(x) = \sqrt{\frac{2m_c^\star}{\hbar^2}(E - E_c(x))}$ is a spatially varying wavevector. Then, the envelope function varies in space as

$$\boxed{C(x) \approx \frac{K'}{\sqrt{k(x)}} e^{\pm i \int_a^x du k(u)},} \tag{24.34}$$

which looks like a plane wave whose wavelength is smoothly varying in space, and the amplitude is decreasing as the wavevector increases. This situation is shown in Fig. 24.8.

The wavevector $k(x) = \sqrt{\frac{2m_c^\star}{\hbar^2}(E - E_c(x))}$ is large for x where the net kinetic energy $E - E_c(x)$ is large; the electron is moving fast at those points. Near the classical turning points where $E - E_c(x) \to 0$, $k(x)$ decreases, the wavelength stretches out, and the electron slows down. The classical analogy of a ball rolling up and down the valley should be clear. Just as in the classical situation, the *probability* of finding the particle is low at x locations where it is moving fast. This feature is captured in the quantum mechanical wavefunction, since the probability density goes as $|C(x)|^2 \propto \frac{1}{k(x)}$ as indicated by the dashed line.

The dynamics of a quantum particle represented by the wavefunction $\psi(x,t)$ is governed by the time-dependent Schrödinger equation

$$[-\frac{\hbar^2}{2m_e}\frac{d^2}{dx^2}+V(x)]\psi(x,t)=i\hbar\frac{\partial}{\partial t}\psi(x,t) \implies j=\frac{\psi^\star\hat{p}\psi-\psi\hat{p}\psi^\star}{2m_e}, \tag{24.35}$$

where the quantum mechanical probability current density j is obtained from the wavefunction. Note that the electron mass is in the denominator, and ψ is the complete wavefunction. Now we recognize the effective mass equation $E_c(-i\nabla)=E_c(x)-\frac{\hbar^2}{2m_c^\star}\frac{d^2}{dx^2}$ which makes the effective mass equation mathematically identical to the free-electron Schrödinger equation:

$$[-\frac{\hbar^2}{2m_c^\star}\frac{d^2}{dx^2}+E_c(x)]C(x,t)=i\hbar\frac{\partial}{\partial t}C(x,t) \implies j=-\frac{i\hbar}{2m_c^\star}[C^\star\frac{\partial C}{\partial x}-C\frac{\partial C^\star}{\partial x}]. \tag{24.36}$$

and the quantum mechanical probability current density j is thus obtained directly from the effective mass parameters, by using the envelope function $C(x)$ and the effective mass $m_c^\star$ instead of the free–electron mass.

Now since the charge current carried by an effective mass state is

$$J=\sum_k j_k = qg_sg_v\sum_k v_g(k)|C(x)|^2, \tag{24.37}$$

we get from Equation 24.34 for a purely right-going WKB state $C(x)=\frac{Ke^{i\int_a^x k(u)du}}{\sqrt{k(x)}}$ in a smoothly varying conduction band profile $E_c(x)$. Because near the band edge the bandstructure is $E_c(k)=E_c(x)+\frac{\hbar^2 k(x)^2}{2m_c^\star}$, the group velocity of the effective mass state is $v_g(k)=\frac{1}{\hbar}\frac{\partial E_c(k)}{\partial k}=\frac{\hbar k(x)}{m_c^\star}$.

The charge current density is then found to be

$$J=qg_sg_v\underbrace{\frac{\hbar k(x)}{m_c^\star}}_{v(x)}\underbrace{|C(x)|^2}_{n(x)}=qg_sg_v\frac{\hbar k(x)}{m_c^\star}\frac{|K|^2}{k(x)}=qg_sg_v\frac{\hbar|K|^2}{m_c^\star}, \tag{24.38}$$

showing how the net current is conserved in transport across a smoothly varying potential profile. This is shown in Fig. 24.9 for an effective mass electron wavepacket in the conduction band of a semiconductor. The envelope carrier density $n(x)=C^\star(x)C(x)=\frac{|K|^2}{k(x)}$ decreases when the group velocity $\frac{\hbar k(x)}{m_c^\star}$ increases, keeping their product $J(x)\sim n(x)v(x)$ constant.

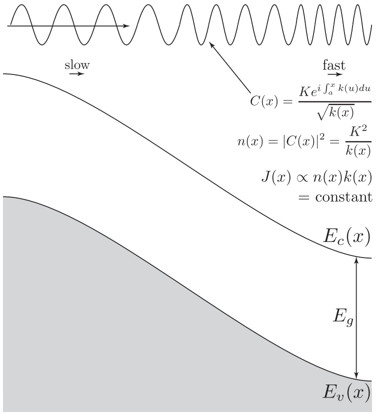
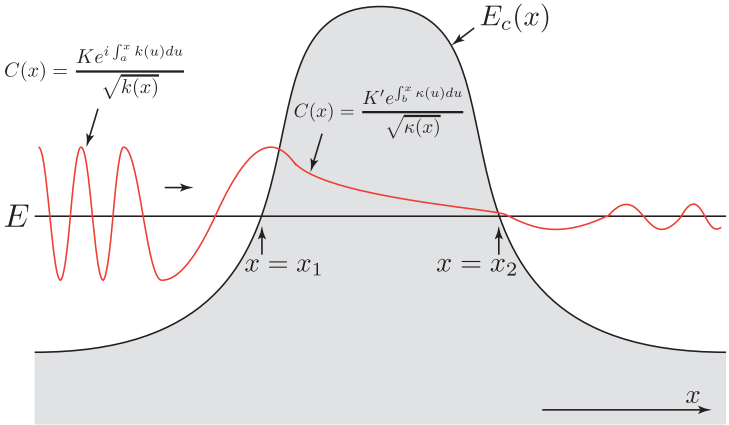

slow
fast
$C(x)=\frac{Ke^{i\int_a^x k(u)du}}{\sqrt{k(x)}}$
$n(x)=|C(x)|^2=\frac{K^2}{k(x)}$
$J(x)\propto n(x)k(x)$
$=$ constant
$E_c(x)$
E_g
$E_v(x)$

Fig. 24.9 Transport of an effective mass electron wavepacket in the conduction band in a smoothly varying potential. The group velocity $v(x) \sim k(x)$ increases as the kinetic energy increases, but the carrier density $n(x)=|C(x)|^2 \sim 1/k(x)$ decreases, keeping the net current $J(x)\sim n(x)v(x)$ constant.

Tunneling transport using WKB: The electron may encounter potential barriers of height larger than its kinetic energy, i.e. $E<V(x)$ as indicated in Fig. 24.10. In such a situation, the term $Q(x)=\kappa(x)=-\frac{2m_c^\star}{\hbar^2}(E-E_c(x))>0$, and the WKB solution of Equation 24.33 is



$C(x) = \frac{Ke^{i\int_a^x k(u)du}}{\sqrt{k(x)}}$
$E_c(x)$
$C(x) = \frac{K'e^{\int_b^x \kappa(u)du}}{\sqrt{\kappa(x)}}$
E
$x = x_1$
$x = x_2$
x

Fig. 24.10 WKB tunneling through a potential barrier.

$$\boxed{C(x) \approx \frac{K}{\sqrt{\kappa(x)}} e^{\pm \int_a^x \kappa(u)du}.} \tag{24.39}$$

Fig. 24.11 Igor Tamm proposed the existence of surface states in semiconductors invoking tunneling of Bloch states into the vacuum. He introduced the concept of phonons. Tamm was awarded the Nobel Prize in Physics in 1958 for his joint discovery of the Cerenkov radiation.

If the classical turning points are $x = x_1$ and $x = x_2$ as indicated in Fig. 24.10, we write the ratio of the coefficients approximately as

$$\frac{C(x_2)}{C(x_1)} \approx \frac{e^{-\int_a^{x_2} \kappa(u)du}}{e^{-\int_a^{x_1} \kappa(u)du}} \approx e^{i\phi} e^{-\int_{x_1}^{x_2} \kappa(u)du}, \tag{24.40}$$

where ϕ is a possible phase factor accumulated in the oscillatory part. Because the tunneling probability is proportional to the square of the wavefunction, we obtain

$$\left|\frac{C(x_2)}{C(x_1)}\right|^2 \approx e^{-2\int_{x_1}^{x_2} \kappa(u)du} \implies \boxed{T_{wkb} \approx e^{-2\int_{x_1}^{x_2} dx \sqrt{\frac{2m_c^\star}{\hbar^2}[E_c(x)-E]}}.} \tag{24.41}$$

This approximate equation holds for tunneling within a single band. It is not exact, but provides a reasonable estimate of the tunneling probability. It may be compared with the exact tunneling probability obtained in the rectangular barrier in Equation 24.2. Using the physical constants, we obtain $T_{wkb} \approx e^{-(\frac{t_b}{0.1\text{ nm}})\sqrt{(\frac{m_c^\star}{m_e})\cdot(\frac{V_0}{1\text{ eV}})}}$. For example, the tunneling probability for a barrier $t_b = 1$ nm thick and uniform barrier height $V_0 \sim 1$ eV, with a conduction band edge effective mass $m_c^\star \sim 0.09 m_e$ is $T_{wkb} \approx e^{-3} \sim 1/20$, a rather high number because of the thin barrier and the light effective mass. The tunneling probability is extremely sensitive to the thickness, barrier height, and the effective mass. This sensitivity is also a reason the tunneling currents can be used to measure these parameters.

24.5 Nordheim supply function

For a large number of electrons, the total tunneling current is found by integrating the product of the number of electrons that meet the criteria for tunneling, with their WKB tunneling probabilities. The last sections focused on the tunneling probabilities. In this section, we discuss the carrier distribution. An important result is the **Nordheim supply function**, first obtained by L. Nordheim. Let us call the direction in which the net tunneling current flows as the z-direction, and the net kinetic energy of carriers in a parabolic free-electron bandstructure $\frac{\hbar^2(k_x^2+k_y^2+k_z^2)}{2m_e} = \frac{\hbar^2(k_x^2+k_y^2)}{2m_e} + E_z$ where $E_z = \frac{\hbar^2 k_z^2}{2m_e}$ is the z-component of the kinetic energy (Fig. 24.12). The supply function is defined as

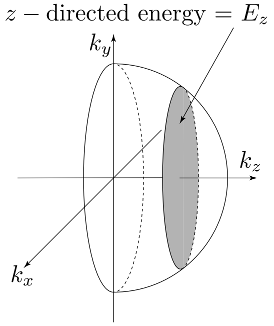


Fig. 24.12 Electron states in the **k**-space which have z-directed kinetic energy E_z lie in the shaded circular slice. The tunneling probability of all electrons on such a circle will be found to be the same in the next section.

$$N(E_z)dE_z = \frac{\#\text{ of electrons with } z\text{-directed energy in } (E_z, E_z + dE_z)}{\text{unit area} \cdot \text{unit time}}. \tag{24.42}$$

The Nordheim supply function $N(E_z)$ for 3D free–electrons with spin degeneracy $g_s = 2$, in units of $1/(\text{cm}^2 \cdot \text{eV} \cdot \text{s})$, is therefore

$$N(E_z)dE_z = \frac{2}{L^3}\sum_{k_x,k_y}\frac{\hbar k_z}{m_e}f(k) = 2\frac{\hbar k_z dk_z}{2\pi m_e}\int\frac{dk_x dk_y}{(2\pi)^2}\frac{1}{1+e^{\left[\frac{\frac{\hbar^2(k_x^2+k_y^2)}{2m_e}-(E_F-E_z)}{k_bT}\right]}}.$$

$$\implies N(E_z) = \frac{4\pi m_e k_b T}{h^3}\ln[1+\exp[\frac{E_F-E_z}{k_bT}]]. \tag{24.43}$$

The generalized Nordheim supply function in d-dimensions and any effective mass, and the tunneling current are[3]

$$\boxed{N_d(E_z) = \frac{N_c^{d-1}}{2\pi\hbar}\cdot F_{\frac{d-3}{2}}(\frac{E_F-E_z}{k_bT})}, \text{ and } \boxed{J_d = q\int dE_z N_d(E_z)T_{wkb}(E_z).} \tag{24.44}$$

24.6 Fowler–Nordheim tunneling

Millikan and Lauritsen discovered in experiments that the current density J drawn from a cold metal into vacuum at an electric field F varied as $J = AF^2\exp[-B/F]$, where A, B are constants, with a negligible temperature dependence. Fowler and Nordheim explained this phenomena of field–emission by realizing that electrons can tunnel out of a metal into vacuum on the application of a large electric field. For the tunneling of electrons from a metal into a semiconductor (not vacuum), consider the energy band diagram shown in Fig. 24.13. The Schottky barrier height is $q\phi_b$. Let us assume an intrinsic semiconductor layer of thickness d_0 followed by a heavily doped contact layer.

The electrons in the metal are the source of the current. We will assume a free-electron bandstructure for the metal, with electrons filled

[3] As always, here N_c^{d-1} is the band-edge DOS in $d-1$ dimensions and $F_j(...)$ is the Fermi–Dirac integral of order j.

up to the Fermi level E_{Fm} from the bottom of the conduction band (of the metal) as shown in Fig. 24.13 schematically. The layer structure figure indicates that electrons are moving in all three directions in the metal. Of these, only electrons with sufficient kinetic energy due to velocity $\hbar k_z/m_e$ pointed *towards* the semiconductor (in the $+z$ direction) have a chance to tunnel into the semiconductor. The z-directed kinetic energy is $E_z = \frac{\hbar^2 k_z^2}{2m_e}$. It is important to realize that electrons with no E_z cannot tunnel, even if they have high energies, say at $E = E_{Fm}$ because they are moving in the wrong direction.

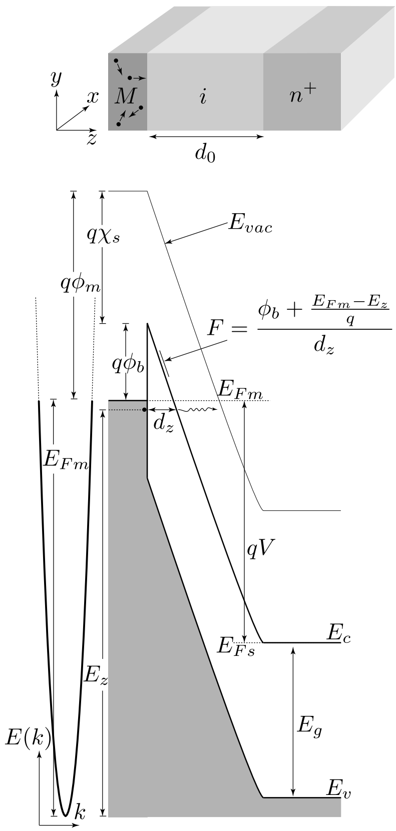


Fig. 24.13 Illustration of Fowler–Nordheim tunneling in a metal-semiconductor junction. The bandstructure of the metal shown schematically on the left is assumed to be parabolic with free–electron mass $E(k) = \frac{\hbar^2 k^2}{2m_e}$, and filled up to E_{Fm}. The right-moving electrons at the metal Fermi level E_{Fm} are incident on the potential barrier of height $q\phi_b$ and thickness $d = \phi_b/F$, and have probability T_{wkb} of tunneling into the semiconductor, where T_{wkb} is evaluated in Equation 24.45. Electrons in the metal at energies far below E_{Fm} see a much higher barrier and thicker barrier. Though for simplicity an intrinsic undoped i-layer is assumed, such a layer is typically lightly doped. This does not change the discussion and results appreciably.

Consider an electron in the metal with the z-directed kinetic energy E_z. From Fig. 24.13, it sees a triangular barrier whose maximum height at the metal/semiconductor interface is $q\phi_b + E_{Fm} - E_z$. It has to tunnel a distance d_z, where $d_z = (\phi_b + \frac{E_{Fm}-E_z}{q})/F$, where $F = (V + \phi_b)/d_0$ is the relation of the applied electric field to the applied voltage. Since the potential barrier shape varies as $(q\phi_b + E_{Fm} - E_z)(1 - \frac{z}{d_z})$, the WKB tunneling probability for the electron of z-directed energy E_z is

$$T_{wkb}(E_z) \approx \exp\left[-\int_0^{d_z} dz \sqrt{\frac{2m_T^\star[(q\phi_b + E_{Fm} - E_z)(1 - \frac{z}{d_z})]}{\hbar^2}}\right]$$

$$= \exp[-\frac{4}{3}\sqrt{\frac{2m_T^\star q\phi_b^3}{\hbar^2}} \cdot \frac{1}{F} \cdot (1 + \frac{E_{Fm} - E_z}{q\phi_b})^{\frac{3}{2}}] \implies$$

$$\boxed{T_{wkb}(E_z) \approx \exp[-\frac{F_0}{F} \cdot (1 + \frac{E_{Fm} - E_z}{q\phi_b})^{\frac{3}{2}}]} \; \& \; \boxed{F_0 = \frac{4}{3}\sqrt{\frac{2m_T^\star q\phi_b^3}{\hbar^2}}.} \quad (24.45)$$

The net current density may now be calculated by summing over all **k**-states while satisfying the Pauli exclusion principle. Assuming the metal has spin degeneracy $g_s = 2$ and valley degeneracy $g_v = 1$, using the fact that the occupation function on the right in the conduction band of the semiconductor depletion region is nearly zero ($f_R \approx 0$) and also using $E(k) = \frac{\hbar^2}{2m_e}(k_x^2 + k_y^2) + E_z$ in the metal, we get

$$J = \frac{qg_sg_v}{L^3}\sum_k [v_{gL}(k)T_{wkb}(E_z)f_L(1 - f_R) - v_{gR}(k)T_{wkb}(E_z)f_R(1 - f_L)]$$

$$\approx \frac{2q}{L^3}\sum_k v_{gL}(k)T_{wkb}(E_z)f_L = \frac{2q}{L^3}\int \frac{dk_xdk_ydk_z}{(\frac{2\pi}{L})^3}\frac{\hbar k_z}{m_e}\frac{T_{wkb}(E_z)}{1 + \exp[\frac{E(k)-E_{Fm}}{k_bT}]}$$

$$= q\int_{E_z=0}^{\infty} dE_z \cdot N_{3d}(E_z) \cdot T_{wkb}(E_z)$$

$$= \frac{qm_ek_bT}{2\pi^2\hbar^3}\int_0^{\infty} dE_z \ln[1 + \exp[\frac{E_{Fm} - E_z}{k_bT}]] \cdot \exp[-\frac{F_0}{F} \cdot (1 + \frac{E_{Fm} - E_z}{q\phi_b})^{\frac{3}{2}}]$$

$$\approx \frac{qm_e}{2\pi^2\hbar^3}\int_0^{E_{Fm}} dE_z(E_{Fm} - E_z)\exp[-\frac{F_0}{F}(1 - \frac{3}{2}\frac{E_{Fm} - E_z}{q\phi_b})]$$

$$\implies \boxed{J_{FN}(F) \approx \frac{m_e}{m_T^\star} \cdot \frac{q^2}{16\pi^2\hbar\phi_b} \cdot F^2 \cdot \exp[-\frac{F_0}{F}] = J_{FN}^0(\frac{F}{F_0})^2\exp[-\frac{F_0}{F}].} \quad (24.46)$$

The exact energy integral includes the weak temperature dependence and may be evaluated numerically[4]. Since the tunneling barrier is the lowest for electrons with the highest E_z, the electrons in the metal with $E_z \approx E_{Fm}$ have appreciable tunneling. For electrons with $E_z << E_{Fm}$, the tunneling probability is exponentially damped, whereas energy states $E_z >> E_{Fm}$ have very small occupation due to the Fermi–Dirac distribution. Also, for $E_z \approx E_{Fm}$, we must have $\frac{E_{Fm}-E_z}{q\phi_b} << 1$ since $E_{Fm} - E_z$ is typically a few k_bT whereas $q\phi_b$ is typically of the order of 1 eV. The last two lines of Equation 24.46 take advantage of these facts to extract an analytical formula for the tunneling current density. With these approximations, $\ln[1 + \frac{E_{Fm}-E_z}{k_bT}] \rightarrow \frac{E_{Fm}-E_z}{k_bT}$ for $E_z < E_{Fm}$ and $\ln[1 + \frac{E_{Fm}-E_z}{k_bT}] \rightarrow 0$ for $E_z > E_{Fm}$. This limits the integral to $E_z < E_{Fm}$, and allows a first order Taylor approximation in the exponential. The final boxed form in Equation 24.46 further retains only the leading term in the analytical integral[5] (see Exercise 24.2).

In the boxed form of Equation 24.46 $J^0_{FN} = \frac{2q^3 m_e \phi_b^2}{9\pi^2\hbar^3}$ is the characteristics Fowler–Nordheim tunneling current density dependent **only on the barrier height**. Since F_0 depends on the barrier height ϕ_b and the tunneling effective mass $m_T^\star$ from Equation 24.45, the Fowler–Nordheim tunneling current density is obtained simply by multiplying J^0_{FN} with $u \cdot \exp[-1/u]$ with $u = F_0/F$.

Consider first electrons with $E_z = E_{Fm}$: for them, $T_{wkb} \approx \exp[-F_0/F]$, which already captures the correct experimental dependence. The WKB tunneling probability for most semiconductors follows this type of $\exp[-F_0/F]$ behavior. The characteristic electric field F_0 depends on the tunneling effective mass of the semiconductor and the Schottky barrier height. When $F << F_0$, T_{wkb} is very small. But as soon as $F \geq F_0$, $T_{wkb} \approx 1$ and tunneling is all-important, as if the barrier did not exist. The characteristic electric field is $F_0 \approx 68\sqrt{\frac{m_c^\star}{m_e}}(\frac{\phi_b}{1.0\text{V}})^{\frac{3}{2}}$ MV/cm, which for $m_c^\star = 0.05 m_e$ and $\phi_b = 0.3$ eV is $F_0 \approx 2.5$ MV/cm.

The field F dependence enters in the **denominator** of the exponential, not the numerator. It is important to realize that the form of this exponential is deceptively non-analytic: it cannot be Taylor expanded around any value of F! This is a signal that the **tunneling probability cannot be obtained from time-independent perturbation theory**. The problem must be treated head-on by solving the Schrödinger equation itself, as we did in the earlier sections of this chapter.

Fowler–Nordheim tunneling is observed in most clean semiconductor junctions as the mechanism for current flow at large reverse-bias voltages, and places a limit on the reverse-blocking voltages of Schottky diodes. Increasing the tunneling barrier height (e.g. by use of heterostructures) can reduce the tunneling current[6]. The natural upper limit of the barrier height for bulk semiconductors is the entire bandgap of the semiconductor, which leads to higher reverse-bias blocking voltages. The junctions that exploit the entire bandgap barrier are pn junctions, which require interband tunneling. This is discussed in the next two sections.

---

[4] It should be noted that $m_T^\star$ in Equations 24.45 and 24.46 is a *tunneling effective mass*. Though for Fowler-Nordheim tunneling $m_T^\star \approx m_c^\star$ is typically used for metal–n-type semiconductor Schottky diodes, accurate $m_T^\star$ can be calculated from bandstructure evaluation methods.

[5] Fowler and Nordheim's original solution of this problem from the reflection and transmission amplitudes of the Schrödinger equation, via asymptotic limits of Airy functions, yields an additional dimensionless prefactor: $T(E_z) \approx \frac{4\sqrt{E_z(q\phi_b+E_{Fm}-E_z)}}{q\phi_b+E_{Fm}} \times \exp[-\frac{F_0}{F} \cdot (1 + \frac{E_{Fm}-E_z}{q\phi_b})^{\frac{3}{2}}]$ because of reflection at the abrupt step in the triangular barrier at the metal-semiconductor junction. But because the exponential is so dominant, we shall use the simpler WKB form. The numerical loss of accuracy by making the WKB approximation is not that significant. See Exercise 24.1.

[6] Fowler–Nordheim tunneling is the mechanism of writing data into **flash memories**. In such a device, electrons tunnel into either a nanoscale semiconductor or metallic island from the channel of a MOSFET, where it is trapped between barriers. The data can be sensed by a threshold voltage shift of the MOSFET. The data is erased by tunneling the electrons out by Fowler–Nordheim tunneling. See Exercise 24.3.

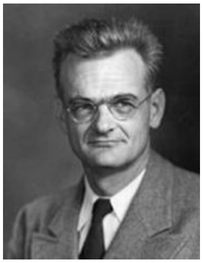


Fig. 24.14 Clarence Zener performed pioneering studies on interband tunneling as a mechanics of electrical breakdown of dielectrics.

24.7 Interband Zener tunneling

At high electric fields in semiconductors, tunneling of electrons can occur from the valence band to the conduction band. Such interband tunneling may only be loosely modeled by the above effective mass formalism, because the effective mass approximation works best for electron transport within the same band. Let us first develop a very approximate method to estimate the interband tunneling currents, and then refine it in successive steps.

The barrier height seen by an electron in the valence band is $E_g + (E_v(k) - E_z)$, since the portion of the kinetic energy perpendicular to the potential barrier $E_\perp = E_v(k) - E_z$ does not contribute to tunneling. Thus, the potential barrier shape is $E_g + (E_v(k) - E_z)(1 - \frac{x}{d})$ where $d = \frac{E_g + (E_v(k) - E_z)}{qF}$, where F is the (constant) electric field. Using the WKB result, the tunneling probability is then

$$T_{wkb}(E_\perp) = \exp\left[-\frac{4}{3}\frac{\sqrt{2m_T^\star(E_g + E_\perp)^3}}{q\hbar F}\right] \approx \exp\left[-\frac{F_0}{F}\right]\exp\left[-\frac{E_\perp}{E_0}\right]$$

$$\text{where } F_0 = \frac{4}{3}\cdot\sqrt{\frac{2m_T^\star E_g^3}{\hbar^2 q^2}} \text{ and } E_0 = \frac{2F}{3F_0}E_g. \quad (24.47)$$

Here F_0 is the characteristic tunneling electric field, and E_0 is the characteristic scale of the perpendicular kinetic energy. We have used the linear approximation $\sqrt{(E_g + E_\perp)^3} \approx E_g^{\frac{3}{2}}(1 + \frac{3}{2}\frac{E_\perp}{E_g})$ in the above, since the perpendicular component is damped strongly as $E_\perp$ increases. This perpendicular component is suppressed in lower dimensions, and completely eliminated in 1D. We next use this tunneling probability to evaluate the net interband tunneling current in pn junction diodes.

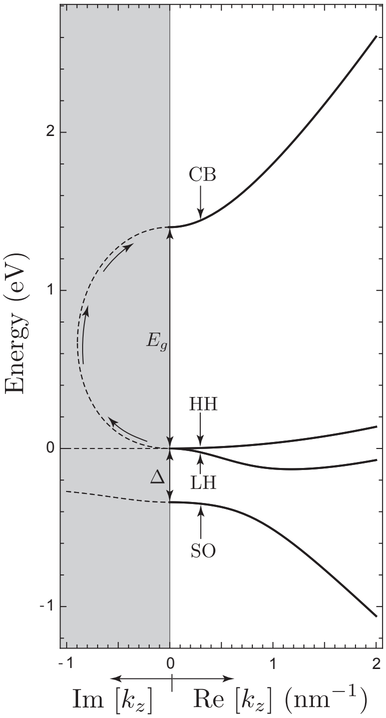


Fig. 24.15 Real bandstructure (right side, thick solid lines) and evanescent energy states (left side, thin dashed lines) inside the bandgap calculated for GaAs from the $\mathbf{k}\cdot\mathbf{p}$ bandstructure (Chapter 12) into the imaginary k-space by replacing $k \to i\kappa$. This is adapted from Figure 12.12. The tunneling amplitude is obtained as a path integral along the dashed line.

24.8 pn tunnel junctions in 1D, 2D, and 3D

Fig. 24.16 (a) shows the energy band diagram for interband tunneling in the pn junction at a large reverse-bias voltage V. Electrons tunnel from the filled valence band states in the p-side on the left to the empty conduction band states in the n-side on the right. The interband tunneling current is the difference of the right going and left going current densities:

$$J_{net} = J_{p\to n} - J_{n\to p}$$

$$= q\frac{g_s g_v}{L^3}\sum_{k\in\Omega_t}\underbrace{(\frac{1}{\hbar}\nabla_{\mathbf{k}}E_v(\mathbf{k})\cdot\hat{z})}_{v_{gz}^p}\cdot T_{wkb}(k)\cdot f_p(1-f_n)$$

$$-q\frac{g_s g_v}{L^3}\sum_{k\in\Omega_t}\underbrace{(\frac{1}{\hbar}\nabla_{\mathbf{k}}E_c(\mathbf{k})\cdot\hat{z})}_{v_{gz}^n}\cdot T_{wkb}(k)\cdot f_n(1-f_p), \quad (24.48)$$

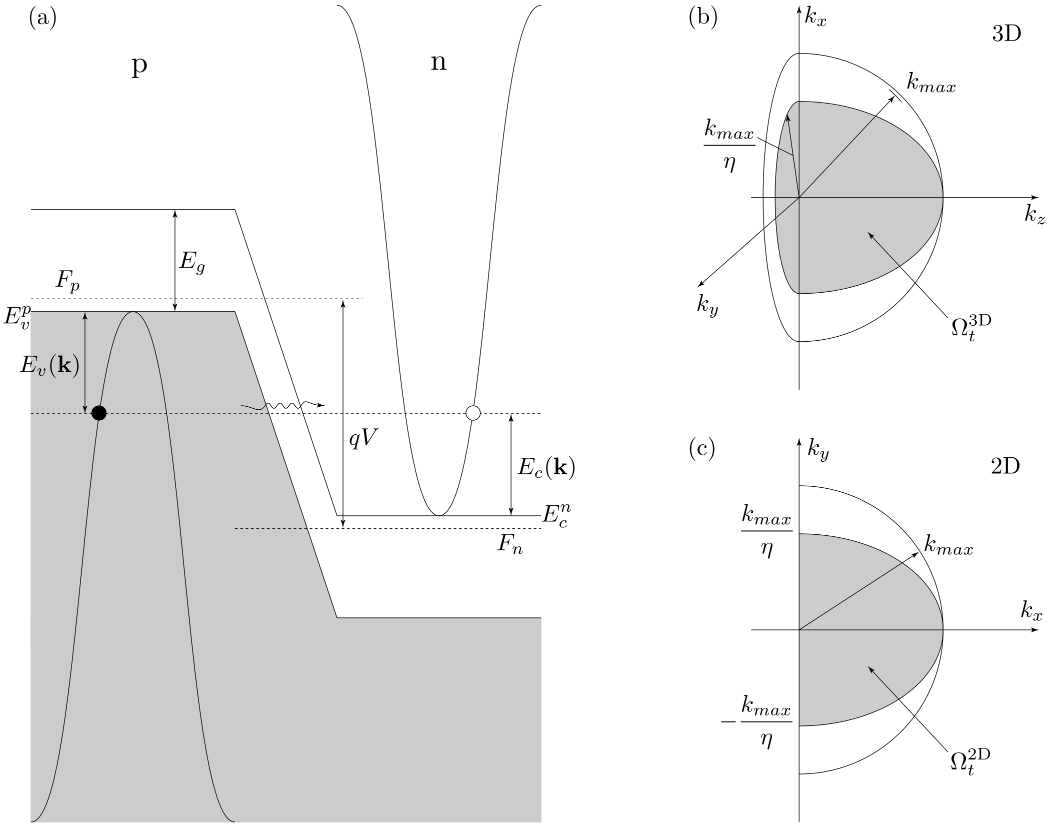


Fig. 24.16 Energy band diagram illustrating interband tunneling.

where the spin and valley degeneracy are assumed to be the same on the respective bands on both sides (this assumption may be relaxed). Here f_p is the Fermi–Dirac occupation function of valence band states in the p-side, and f_n is the corresponding occupation function for the conduction band states in the n-side at the same energy. Because tunneling is a reciprocal process, the WKB tunneling probability is identical for electrons tunneling from $p \to n$ or from $n \to p$. The sums run over all k-states denoted by a k-space "volume" Ω_t for which tunneling is allowed, as is discussed shortly.

To evaluate the above expression for interband tunneling current further, it is necessary to use the valence bandstructure $E_v(\mathbf{k})$ of the p-side and the conduction bandstructure $E_c(\mathbf{k})$ of the n-side. The bandstructures are also necessary to identify the k-values on the p- and n-sides that should be included in the sums. From Fig. 24.16 (a), we will consider the case of elastic tunneling, in which the energy of the electron is conserved in the tunneling process. The perpendicular component of the momentum is conserved since there is no force acting in that direction.

3D pn junction: For transport in the z-direction, $\hbar k_{xn} = \hbar k_{xp}$ and $\hbar k_{yn} = \hbar k_{yp}$. Then[7], energy and momentum conservation require

$$E_v^p - E_v(\mathbf{k}) = E_c^n + E_c(\mathbf{k}), \tag{24.49}$$

which leads to

$$\begin{gathered} E_v^p - \frac{\hbar^2}{2m_v^\star}(k_{xp}^2 + k_{yp}^2 + k_{zp}^2) = E_c^n + \frac{\hbar^2}{2m_c^\star}(k_{xn}^2 + k_{yn}^2 + k_{zn}^2) \text{ and} \\ E_v^p - E_c^n = \frac{\hbar^2}{2m_v^\star}(k_{xp}^2 + k_{yp}^2 + k_{zp}^2) + \frac{\hbar^2}{2m_c^\star}(k_{xn}^2 + k_{yn}^2 + k_{zn}^2), \end{gathered} \tag{24.50}$$

which from Fig. 24.16 gives

$$\underbrace{qV + \zeta_n + \zeta_p}_{qV'} = \frac{\hbar^2}{2}\underbrace{(\frac{1}{m_c^\star} + \frac{1}{m_v^\star})}_{1/m_r^\star}(k_x^2 + k_y^2) + \frac{\hbar^2 k_{zp}^2}{2m_v^\star} + \frac{\hbar^2 k_{zn}^2}{2m_c^\star}.$$

The applied voltage splits the quasi-Fermi levels to $F_p - F_n = qV$. The doping densities on the p-side and n-side determine the location of the respective band edges, denoted by $\zeta_p = E_v^p - F_p$ and $\zeta_n = F_n - E_c^n$. We first consider $\zeta_p = \zeta_n = 0$ when the quasi-Fermi levels are flush with the band edges, and hence $qV' = qV + \zeta_n + \zeta_p = qV$. This also forbids tunneling from $n \to p$ for small forward-bias voltages[8], so we consider only tunneling from $p \to n$.

The range $\Omega_{t,p}^{3D}$ of k states in the valence band for tunneling of electrons from $p \to n$ is obtained immediately from Equation 24.51. The momentum in the direction of tunneling is not required to be conserved because of the presence of a large force in that direction. However, after an electron tunnels through the bandgap, it must emerge on the n-side. This requires $k_{zn}^2 \geq 0$, which directly yields the range of k-states for tunneling both ways from $p \to n$ and $n \to p$:

$$\begin{gathered} \frac{\hbar^2 k_\perp^2}{2m_r^\star} + \frac{\hbar^2 k_{zp}^2}{2m_v^\star} \leq qV \implies k_{zp}^2 + \underbrace{\frac{m_v^\star}{m_r^\star}}_{\eta^2} k_\perp^2 \leq \underbrace{\frac{2m_v^\star qV}{\hbar^2}}_{k_{max}^2}, \\ \implies \Omega_{t,p}^{3D} : \boxed{k_{zp}^2 + \eta^2 k_\perp^2 \leq k_{max}^2,} \text{ and } \Omega_{t,n}^{3D} : \boxed{k_{zn}^2 + (1 - \eta^2) k_\perp^2 \leq k_{max}^2.} \end{gathered} \tag{24.51}$$

The k-space volume $\Omega_{t,p}^{3D}$ that consists of the states that can tunnel from $p \to n$ is shown as the 3D ellipsoid in Fig. 24.16 (b). Because they must satisfy $k_{zp}^2 + \eta^2 k_\perp^2 \leq k_{max}^2$, the radius in the $k_x - k_y$ plane is k_{max}/η. Equations 24.48 and 24.47 then give the tunneling current

[7] Though we specifically consider parabolic valence and conduction bandstructures, the same method applies in the general case.

[8] These restrictions are relaxed later. The case of very heavy degenerate doping when $F_p < E_v^p$ and $F_n > E_c^n$, and therefore $\zeta_p > 0$ and $\zeta_n > 0$, leads to the interesting phenomena of negative differential resistance in Esaki diodes.

from $p \to n$ as the integral over the states in the ellipsoidal volume:

$$J_T^{3D} \approx J_{p\to n} = \frac{q g_s g_v}{L^3} \sum_{k\in\Omega^{3D}_{t,p}} \frac{\hbar k_{zp}}{m_v^\star} T_{wkb}(k) \underbrace{f_p(k)[1-f_n(k)]}_{\approx 1}$$

$$= \frac{q g_s g_v}{(2\pi)^3} e^{-\frac{F_0}{F}} \int_{k_\perp=0}^{k_{max}/\eta} dk_\perp \cdot 2\pi k_\perp \cdot e^{-\frac{\hbar^2 k_\perp^2}{2m_v^\star \bar{E}}} \cdot \int_{k_{zp}=0}^{\sqrt{k_{max}^2-\eta^2 k_\perp^2}} dk_{zp} \cdot \frac{\hbar k_{zp}}{m_v^\star}$$

$$\implies J_T^{3D} \approx \frac{q^2 g_s g_v}{2\pi\hbar} \cdot \left(\frac{m_v^\star E_g}{3\pi\hbar^2} \cdot \frac{F_0}{F}\right) \cdot e^{-\frac{F_0}{F}} \cdot V \cdot \underbrace{[1-\frac{V_0}{V}(1-e^{-\frac{V}{V_0}})]}_{\approx 1 \text{ for } V>>V_0}$$

$$\implies \boxed{J_T^{3D} \approx \frac{q^2 g_s g_v}{2\pi\hbar} \cdot \left(\frac{m_v^\star E_g}{3\pi\hbar^2} \cdot \frac{F_0}{F}\right) \cdot e^{-\frac{F_0}{F}} \cdot V.} \tag{24.52}$$

In the above, $qV_0 = \eta^2 \bar{E}$ is a characteristic lateral energy scale. Since $\bar{E} = \frac{2F}{3F_0} E_g$, and typical fields $F << F_0$, this energy is a small fraction of the energy bandgap and $V >> V_0$ is satisfied easily at large reverse-bias voltage. In the above we have assumed the occupation functions to be $f_p(k) = 1$ and $f_n(k) = 0$. This approximation enables the analytical evaluation, and need not be made in a numerical calculation. The analytical formula is seen to be a product of the quantum of conductance, the voltage and tunneling probability, and the term in the brackets is an effective inverse area. Fig. 24.17 (a) shows numerical values at $V = 0.5$ volt for various bandgaps and effective masses.

Note the very strong dependence of the Zener tunneling current density on both the energy bandgap E_g of the semiconductor, and the junction electric field F. It must be realized that the high electric fields are typically not entirely due to an externally applied voltage. A large part of the field in realistic Zener tunneling diodes is created in a pn junction by heavy doping of donors and acceptors. A typical rule of thumb is that the interband tunneling distance is a few nm, and the doping as high as feasible in the semiconductor, approaching the solubility limits that are around $10^{20}/\text{cm}^3$ for most semiconductors.

2D pn junction: For a 2D pn junction in which transport occurs in the x-direction, the tunneling probability is $T_{wkb}(k_y) = \exp[-F_0/F] \cdot \exp[-\hbar^2 k_y^2/(2m_v^\star \bar{E})]$. Energy and lateral momentum conservation relations give the area of tunneling k-states

$$\Omega^{2d}_{t,p} : \boxed{k_{xp}^2 + \eta^2 k_y^2 \leq k_{max}^2} \text{ and } \Omega^{2d}_{t,n} : \boxed{k_{xn}^2 + (1-\eta^2) k_y^2 \leq k_{max}^2.} \tag{24.53}$$

Fig. 24.16(c) indicates the area as an ellipse. Using the same nota-

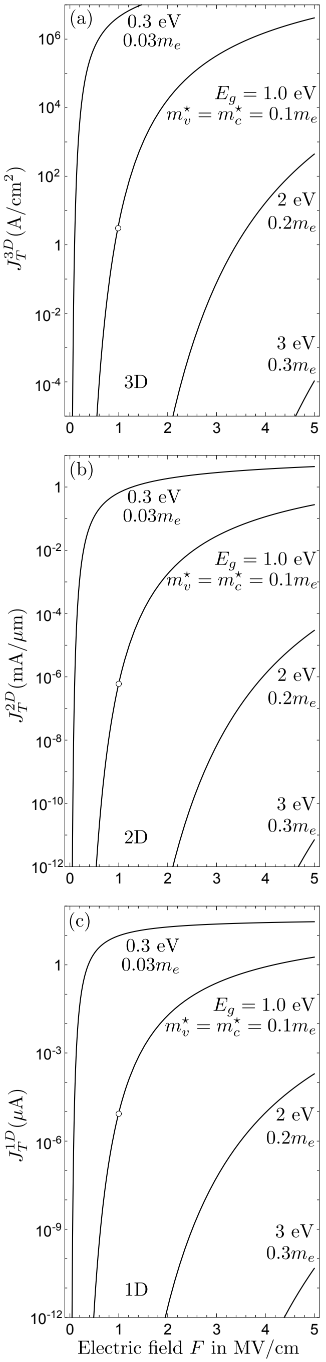


Fig. 24.17 Tunneling currents in 3D, 2D, and 1D pn junctions at reverse-bias of $V = 0.5$ V of various bandgaps and effective masses.

tions as for 3D, the 2D tunneling current from $p \to n$ is then

$$\begin{aligned} J_T^{2D} &\approx J_{p\to n} \approx \frac{q g_s g_v}{L^2} \sum_{k \in \Omega_{t,p}^{2D}} \frac{\hbar k_{xp}}{m_v^\star} T_{wkb}(k) \\ &= \frac{q g_s g_v}{(2\pi)^2} \int_{k_y=-k_{max}/\eta}^{+k_{max}/\eta} dk_y \cdot e^{-\frac{\hbar^2 k_y^2}{2 m_v^\star \bar{E}}} \cdot \int_{k_x=0}^{\sqrt{k_{max}^2 - \eta^2 k_y^2}} dk_{xp} \cdot \frac{\hbar k_{xp}}{m_v^\star} \\ J_T^{2D} &\approx \frac{q^2 g_s g_v}{2\pi\hbar} \cdot \Big(\frac{m_T^\star E_g}{3\pi\hbar^2} \cdot \frac{F}{F_0}\Big)^{\frac{1}{2}} \cdot e^{-\frac{F_0}{F}} \cdot V \times \\ &\underbrace{\Big[\Big(1 - \frac{V_0}{2\sqrt{\pi}V}\Big) \cdot \mathrm{Erf}\Big(\sqrt{\frac{V}{V_0}}\Big) + \sqrt{\frac{V}{\pi V_0}} e^{-\frac{V}{V_0}}\Big]}_{\approx 1 \text{ for } V >> V_0} \\ &\implies \boxed{J_T^{2D} \approx \frac{q^2 g_s g_v}{2\pi\hbar} \cdot \Big(\frac{m_T^\star E_g}{3\pi\hbar^2} \cdot \frac{F}{F_0}\Big)^{\frac{1}{2}} \cdot e^{-\frac{F_0}{F}} \cdot V.} \end{aligned} \tag{24.54}$$

Fig. 24.17 (b) shows numerical values of the 2D tunneling current per unit width at $V = 0.5$ volt.

1D pn junction: Since for 1D tunneling transport all momentum is in the direction of tunneling, $T_{wkb} = e^{-F_0/F}$. Energy conservation requires $E_v^p - \frac{\hbar^2 k_v^2}{2m_v^\star} = E_c^n + \frac{\hbar^2 k_c^2}{2m_c^\star}$, which means $\frac{\hbar^2 k_v^2}{2m_v^\star} + \frac{\hbar^2 k_c^2}{2m_c^\star} = qV$. The interband tunneling current of a 1D pn junction has an exact analytical solution at finite temperatures. To show that, we retain the full Fermi–Dirac occupation functions and write Equation 24.48 as

$$\begin{aligned} J_T^{1D} &= J_{p\to n} - J_{n\to p} \\ &= q \frac{g_s g_v}{L} e^{-\frac{F_0}{F}} \Big[\sum_{k_v} \frac{\hbar k_v}{m_v^\star} \cdot f_p(1 - f_n) - \sum_{k_c} \frac{\hbar k_c}{m_c^\star} \cdot f_n(1 - f_p)\Big] \\ &= \frac{q g_s g_v k_b T}{2\pi\hbar} \Big[\int_0^v du_v \frac{1}{1 + \exp[-u_v]} \Big(1 - \frac{1}{1 + \exp[-u_v + v]}\Big) \\ &\quad - \int_0^v du_c \frac{1}{1 + \exp[u_c]} \Big(1 - \frac{1}{1 + \exp[u_c - v]}\Big)\Big] \\ &\implies \boxed{J_T^{1D} = \frac{q g_s g_v k_b T}{2\pi\hbar} \cdot e^{-\frac{F_0}{F}} \cdot \ln\Big[\cosh^2\Big[\frac{qV}{2k_b T}\Big]\Big].} \end{aligned} \tag{24.55}$$

In the above, we have used the notation $u_v = \frac{\hbar^2 k_v^2}{2m_v^\star k_b T}$, $u_c = \frac{\hbar^2 k_c^2}{2m_c^\star k_b T}$, and $v = \frac{qV}{k_b T}$. For voltages $qV >> k_b T$ the above exact form reduces to

$$J_T^{1D} \approx \frac{q^2 g_s g_v}{2\pi\hbar} \cdot e^{-\frac{F_0}{F}} \cdot V \cdot \Big(\underbrace{1 - \frac{k_b T \ln 4}{qV}}_{\approx 1 \text{ for } qV >> k_b T}\Big) \implies \boxed{J_T^{1D} \approx \frac{q^2 g_s g_v}{2\pi\hbar} \cdot e^{-\frac{F_0}{F}} \cdot V.} \tag{24.56}$$

Fig. 24.17 (c) shows numerical values of the 1D tunneling current at $V = 0.5$ Volt.

Comparing the final boxed versions of Equation 24.52, 24.54, and 24.56, the interband tunneling current density in a d-dimensional pn junction at reverse-bias V is given by the generalized expression

$$\boxed{J_T^d \approx \frac{q^2 g_s g_v}{2\pi\hbar} \times \left(\frac{m_v^\star E_g}{3\pi\hbar^2} \times \frac{F}{F_0}\right)^{\frac{d-1}{2}} \times e^{-\frac{F_0}{F}} \times V,} \tag{24.57}$$

which implies the tunneling current density varies as $F \cdot e^{-F_0/F}$ in 3D, as $\sqrt{F} \cdot e^{-F_0/F}$ in 2D, and as $e^{-F_0/F}$ in 1D. As Figs. 24.17 (a, b, and c) show, the current densities depend very strongly on the junction electric field, and the energy bandgaps. The circles indicate the natural scales for $E_g = 1.0$ eV, $m_v^\star = m_c^\star = 0.1 m_e$ at $V = 0.5$ V and $F = 1$ MV/cm. The tunneling current densities are $J_T^{3D} \approx 4.2$ A/cm^2, $J_T^{2D} \approx 6.2 \times 10^{-7}$ mA/μm, and $J_T^{1D} \approx 9.2 \times 10^{-6}$ μA. As a calibration, these natural scales are roughly 10^6 times smaller than the on-current densities of typical high-performance bipolar transistors ($J_{on} \approx 10^6$ A/cm^2), or field-effect transistors operating using 2D channels ($J_{on} \approx 1$ mA/μm), or 1D channels ($J_{on} \approx 10$ μA). Note also that the explicit dependence on the junction electric field F is much stronger than the voltage, though the relation between voltage and field of a p-i-n diode can be used in this formulation directly.

24.9 NDR, backward diodes

A most remarkable effect of **negative differential resistance** or NDR was observed by Leo Esaki (Fig. 24.18) while measuring the $J - V$ characteristics of heavily doped Ge pn diodes. In addition to the reverse-biased tunneling current given by Equation 24.52, a peak was seen in the forward-bias, followed by a valley, before transitioning to the characteristics diffusion-current dominated $J = J_0(e^{qV/k_bT} - 1)$ turn-on of the pn diode (which was discussed in Chapter 18). This type of heavily doped diode is called an Esaki diode. Fig. 24.19 shows the rich variation of the $J - V$ characteristics of pn diodes with increasing doping density of the p- and n-sides of the junction. Let us discuss them step by step qualitatively first, and then derive a quantitative expression for both the reverse-bias and forward-bias tunneling currents including the negative differential resistance.

Fig. 24.18 Leo Esaki discovered the phenomenon of negative differential resistance in heavily doped semiconductor diodes, and explained it by invoking quantum tunneling and the Pauli exclusion principle. With Raphael Tsu, he proposed the concepts of semiconductor superlattices. Esaki was awarded the 1973 Nobel Prize in Physics with Giaever and Josephson, who discovered tunneling phenomena in superconductors.

(a) "Standard" diode: Fig. 24.19 (a) indicates a lightly doped pn junction, with nominally an undoped i- region between them. A p-i-n diode of this type behaves nearly like an 'ideal' diode as discussed in Chapter 18, following the ideal diode equation in the forward-bias condition, turning on at $V_{bi} \approx (k_bT/q)\ln(N_A N_D/n_i^2)$. At very large reverse-bias voltages, the pair-production process of impact ionization causes avalanche breakdown. This many-particle process, distinct

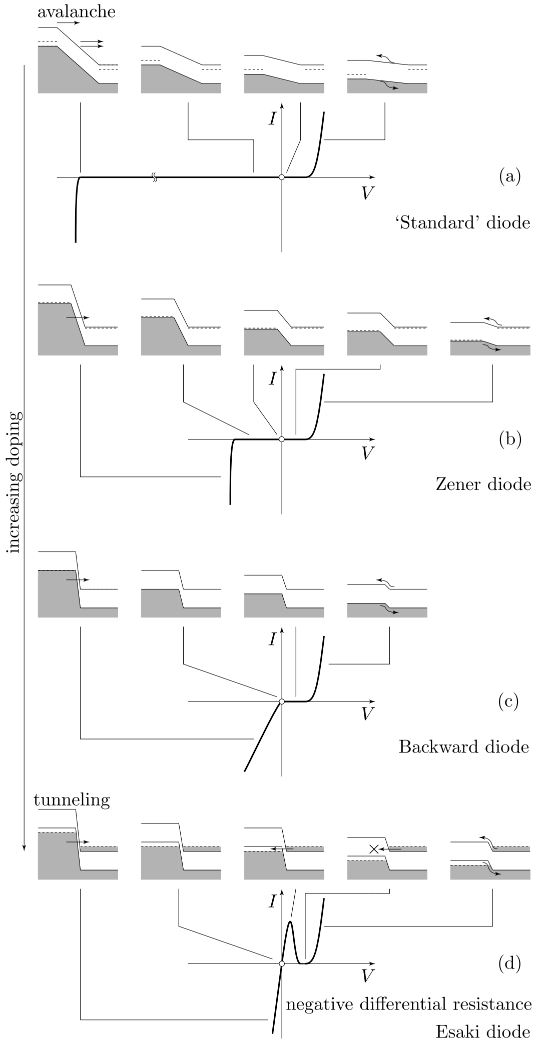

avalanche
increasing doping
I
V
(a)
'Standard' diode
I
V
(b)
Zener diode
I
V
(c)
Backward diode
tunneling
I
V
(d)
negative differential resistance
Esaki diode

Fig. 24.19 Tunneling in pn diodes: evolution of characteristics with increasing doping density.

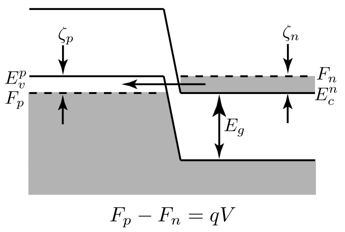

ζ_p
ζ_n
E_v^p
F_p
F_n
E_c^n
E_g
$F_p - F_n = qV$

Fig. 24.20 Energy band diagram and energies in interband tunneling.

from tunneling, is discussed in Section 24.15 later in this chapter. We point out in advance that the impact ionization coefficient α measured in 1/length depends on the electric field as $\alpha(F) = Ae^{-B/F}$, which is similar to the dependence of the tunneling probability. Significant impact ionization occurs at a smaller field than significant interband tunneling and therefore is the primary mechanism for reverse-breakdown of lightly doped pn and p-i-n diodes.

(b) Zener diode: The depletion region thickness shrinks with increasing doping of the p- and n-regions of the diode. Fig. 24.19 (b) shows a qualitatively unchanged forward-bias turn-on as Fig. 24.19 (a), but a much smaller reverse-bias "breakdown". This is because at a large reverse-bias voltage, there is significant interband tunneling.

Diodes that exhibit this type of abrupt increase in the reverse current due to interband tunneling are called Zener diodes.

(c) Backward diode: When the doping densities on both sides of the junction are just about degenerate, approximately equal to the band-edge density of states $N_D \approx N_c$ and $N_A \approx N_v$, the Fermi levels become flush with the band-edges: $E_F = E_v$ on the p-side and $E_F = E_c$ on the n-side. This is shown in Fig. 24.19 (c). The depletion region is thin, and a small reverse-bias aligns the filled valence band states with the empty conduction band states. Therefore the diode starts conducting at small reverse-bias voltages. But for forward-bias voltages $V << V_{bi}$, the forward-bias current remains low since the filled states in the valence band on the p-side point to the bandgap on the n-side. Thus, this diode conducts more current in the reverse-bias than the forward-bias, and is called a **backward diode**. The reverse-bias conduction is due to interband tunneling, with current density given by Equation 24.57. See Exercise 24.7 for the interesting history of backward diodes.

(d) Esaki diode: Fig. 24.19 (d) shows that when the doping densities are so high that the Fermi levels are inside the bands, a small tunneling window is opened at forward-bias voltages in which electrons can tunnel from the conduction band of the n-side to the empty valence band states on the p-side. As the forward-bias voltage increases, this window first increases, reaches a maximum, and then gradually the filled states in the conduction band point into the bandgap on the other side, thus leading to a peak and a valley. Such diodes are called Esaki diodes. At larger forward-bias voltages, the standard diffusion term takes over. Therefore in the voltages in which $dJ/dV < 0$, the Esaki diode exhibits a negative differential conductance, or equivalently a negative differential resistance (NDR). Biasing the diode in the NDR region produces high frequency oscillations due to the instability (e.g. multiple possible voltages for the same current). Additionally, unlike all other diodes, for small voltages around $V = 0$, the Esaki diode behaves like a resistor following Ohm's law. This property makes it very useful for making interband tunnel junctions which are used in several electronic and photonic devices. A prime example is in multijunction solar cells, which are discussed in Chapter 28.

Fig. 24.21 shows the numerically computed tunneling current densities of 3D pn junctions of various doping densities. We will now obtain the expression that is used for this calculation of the tunneling current in both the forward and reverse-bias voltages, including the backward-diode or the NDR regimes. This does not require new concepts: we just need to allow for generalized doping and location of the Fermi levels (see Fig. 24.20), and remove the restriction $\zeta_p = \zeta_n = 0$ we had placed earlier in evaluating the current in Equation 24.52. Furthermore, because to understand NDR we can no longer neglect the $n \to p$ tunneling component, we evaluate the entire tunneling current

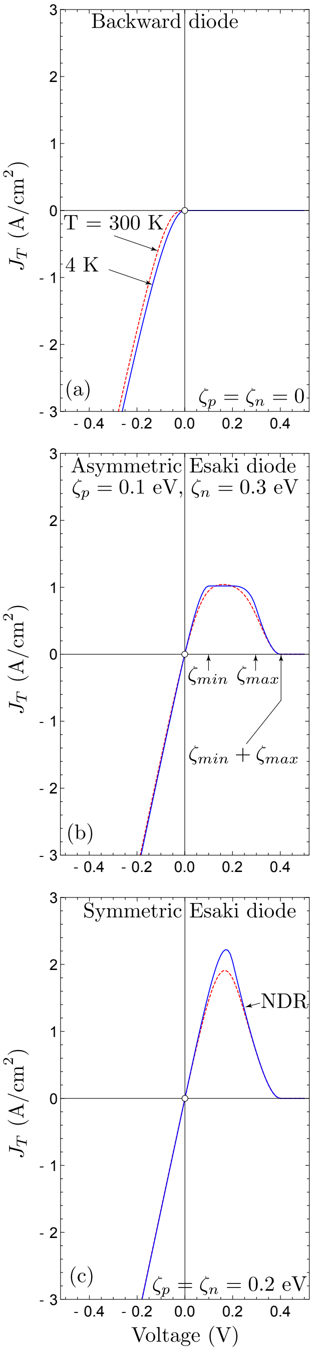


Fig. 24.21 Calculated 300 K and 4 K tunneling current densities in 3D of (a) a backward diode and (b, c) Esaki diodes exhibiting negative differential resistance in forward-bias. The calculation is for $E_g = 1.0$ eV, $m_v^\star = m_c^\star = 0.2 m_e$, $g_s = 2$, $g_v = 1$, and $F = 1$ MV/cm.

as the difference:

$$
\begin{gathered}
J_T = \frac{q g_s g_v}{L^d}\Big[\sum_{k_v \in \Omega^{3D}_{t,p}} \frac{\hbar k_{vz}}{m_v^\star} e^{-\frac{F_0}{F}} e^{-\frac{E_{v\perp}}{E_0}} f_p(1-f_n) \\
- \sum_{k_c \in \Omega^{3D}_{t,n}} \frac{\hbar k_{cz}}{m_c^\star} e^{-\frac{F_0}{F}} e^{-\frac{E_{c\perp}}{E_0}} f_n(1-f_p)\Big], \text{ where} \\
f_p = \frac{1}{1+\exp\big[\frac{(E_v^p - \frac{\hbar^2 k_v^2}{2m_v^\star}) - F_p}{k_b T}\big]}, \text{ and } f_n = \frac{1}{1+\exp\big[\frac{(E_c^n + \frac{\hbar^2 k_c^2}{2m_c^\star}) - F_n}{k_b T}\big]}, \\
\text{where } E_v^p - \frac{\hbar^2 k_v^2}{2m_v^\star} = E_c^n + \frac{\hbar^2 k_c^2}{2m_c^\star} \text{ for elastic tunneling,} \\
k_v^2 = \underbrace{(k_{vx}^2 + k_{vy}^2)}_{k_{v\perp}^2} + k_{vz}^2 \text{ and } k_v^2 = \underbrace{(k_{cx}^2 + k_{cy}^2)}_{k_{c\perp}^2} + k_{cz}^2, \\
\text{with } E_{av} = \frac{\hbar^2 k_{v\perp}^2}{2m_r^\star}, \text{ whereby } E_{v\perp} = \frac{m_r^\star}{m_v^\star} E_{av} \text{ and } E_{c\perp} = \frac{m_r^\star}{m_c^\star} E_{av}, \text{ we get} \\
qV + \zeta_p + \zeta_n = E_{av} + E_{vz} + E_{cz} \text{ from energy conservation, and} \\
J_T^{3D} = \frac{q g_s g_v m_r^\star e^{-\frac{F_0}{F}}}{(2\pi)^2 \hbar^3} \times \\
\Big[\int_0^{qV+\zeta_p+\zeta_n} dE_{av} \cdot e^{-\frac{m_r^\star}{m_v^\star}\frac{E_{av}}{E_0}} \cdot \int_0^{qV+\zeta_p+\zeta_n-E_{av}} dE_{vz} \cdot f_p(1-f_n) - \\
\int_0^{qV+\zeta_p+\zeta_n} dE_{av} \cdot e^{-\frac{m_r^\star}{m_c^\star}\frac{E_{av}}{E_0}} \cdot \int_0^{qV+\zeta_p+\zeta_n-E_{av}} dE_{cz} \cdot f_n(1-f_p)\Big].
\end{gathered}
\tag{24.58}
$$

Though the formulation of Equation 24.58 can be simplified into analytical forms for specific cases (e.g. for $T \to 0$ K), it is in a general form well-suited for numerical evaluation. It is used to calculate the diode characteristics shown in Fig. 24.21 for (a) a backward diode with $\zeta_p = \zeta_n = 0$, (b) an asymmetric Esaki diode with $\zeta_p = 0.1$ eV and $\zeta_n = 0.3$ eV, and (c) for a symmetric Esaki diode with $\zeta_p = \zeta_n = 0.2$ eV. Note the backward diode does not show NDR, but the asymmetric and symmetric Esaki diodes exhibit strong NDRs in forward-bias, with very weak temperature dependence. Note the characteristic voltages $\zeta_{min} = \text{Min}(\zeta_p, \zeta_n)$ and $\zeta_{max} = \text{Max}(\zeta_p, \zeta_n)$ and how they appear in the Esaki diodes and control the peak height and width of the NDR peak. It should be clear from the energy band diagram in Figures 24.19 and 24.20 that for forward-bias voltages greater than $\zeta_p + \zeta_n$ no tunneling current is possible. At forward-bias voltage larger than the built-in voltage $V_{bi} = (k_b T/q)\ln(n_n p_p / n_i^2)$ where n_n and p_p are the majority carrier concentrations of the n- and p-sides, the normal diffusion current dominates by minority carrier injection[9].

[9] In realistic Esaki diodes, the current does not completely go to zero in the voltage window $\zeta_p + \zeta_n \leq qV \leq qV_{bi}$. This is because of defect or trap-assisted tunneling, and also due to the fact that the band-edges are not perfect, but have tail states extending into the gap. These states are either due to defects or heavy doping. The resulting background current limits the ratio of the peak tunneling current and the valley current of the NDR portion. The peak-to-valley-current-ratio (PVCR) of Esaki diodes is a metric of the maximum output power one can extract out of oscillators made from such diodes. PVCRs range from $1-100$ at room temperature depending on the types of diodes.

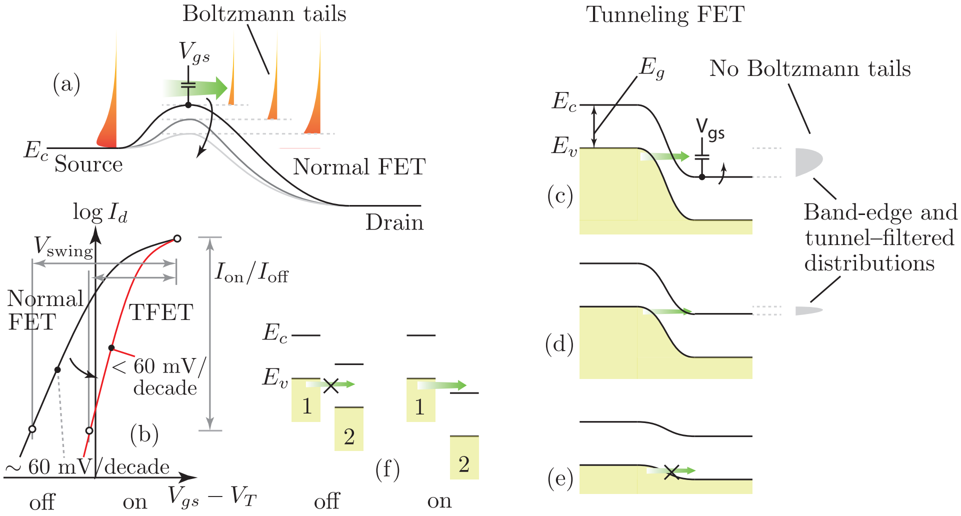

Boltzmann tails
V_{gs}
(a)
E_c
Source
Normal FET
Drain
$\log I_d$
V_{swing}
$I_{\text{on}}/I_{\text{off}}$
Normal FET
TFET
< 60 mV/decade
(b)
~ 60 mV/decade
off
on
$V_{gs} - V_T$
E_c
E_v
1
2
(f)
off
on
Tunneling FET
E_g
No Boltzmann tails
E_c
E_v
$\mathrm{V_{gs}}$
(c)
Band-edge and tunnel–filtered distributions
(d)
(e)

Fig. 24.22 Standard FETs vs. tunneling FETs (**TFETs**).

24.10 Tunneling FETs

In the discussion of field-effect transistors in Chapter 19, it was realized that the sharpest switching slope possible for transport over the potential barrier in the same band is $(k_b T/q)\ln(10)$, which at 300 K is about 60 mV/decade change in current due to a change in the gate voltage. This is because of the Boltzmann tail of high-energy states in the band that are occupied according to the Fermi–Dirac distribution in the band. Fig. 24.22 (a, b) schematically show the energy band diagram and the transfer characteristics of a normal FET operating in the ideal condition of 60 mV/decade.

For low-power applications, a sharper subthreshold slope is desired. This is potentially achieved by using interband, or Zener tunneling as indicated in the energy band diagrams of Figures 24.22 (c, d, and e). Since the DOS in the semiconductor energy bandgap is ideally zero, the high-energy Boltzmann-tail of states is cut off by the gap, resulting in a sharper slope. This is indeed experimentally observed: the resulting subthreshold slope can be substantially lower than 60 mV/decade as schematically shown in Fig. 24.22 (b). As a result, a lower off-state leakage current and a smaller voltage swing is needed, significantly lowering the power consumption of digital circuits.

The first challenge faced by such tunneling FETs (or **TFETs**) is a lower on-state current because of low tunneling probabilities. To increase the interband tunneling current densities, various types of heterostructures with favorable band-alignments are currently under investigation. These include staggered and broken gap alignments, and van der Waals heterojunctions between layered semiconductor materials. As indicated schematically in Fig. 24.22 (f), the primary scheme in such structures is to use the natural band alignment to reduce the tunneling barrier height as far as feasible, while still maintaining a low off-state current. Various heterodimensional structures are being investigated for this purpose, such as tunneling from 3D to 2D states, using combinations of extended and quantized states in semiconductors and their junctions with other materials.

The second challenge is the extreme sensitivity of tunneling to defects and traps that introduce energy levels inside the bandgap of the tunneling barrier region. The third is the broadening of the band-edges via band-tail states. The involvement of these states can increase the on-state current, but they also increase the off-state currents which defeat the goal of low-power operation. Thus TFETs hold much promise, but also demand stringent control and near elimination of defects states and band tails for efficient operation.

24.11 Resonant tunneling

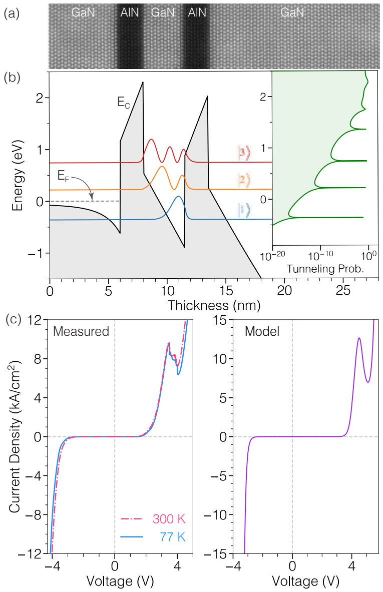


Fig. 24.23 A GaN/**AlN**/GaN/**AlN**/GaN double-barrier resonant tunneling diode. (a) The atomic image of the heterostructure grown by molecular beam epitaxy. (b) The computed energy band diagram and the energy-resolved transmission probability showing resonances. (c) The measured and computed current-voltage characteristics showing negative differential resistance at both 300 K and 77 K. Kindly shared by Dr. Jimy Encomendero, Cornell University.

A rather remarkable consequence of quantum tunneling is exhibited by semiconductor resonant tunneling diodes (**RTDs**). Consider the semiconductor heterostructure shown in Fig. 24.23 (a), with the corresponding energy band diagram in Fig. 24.23 (b). The presence of a single barrier reduces the conductance compared to when there is no barrier due to partial reflection and transmission. The presence of a second barrier should reduce the conductance further – but for certain electron energies, the transmission can become unity. This means that the conductance can be *higher* for two barriers than for one, and this is precisely what is observed in a RTD!

The transmission probability $|T|^2$ across the barriers becomes unity when the incident electron energy is resonant with that of the quasi-bound states of the quantum well between the two barriers. The states are shown as $|1\rangle$, $|2\rangle$, and $|3\rangle$ in Fig. 24.23 (b), with the calculated tunneling probability of the electron across the barriers from the left to the right side. Note that this is tunneling within the same band. Ordinary WKB approximation fails to account for the resonances: instead, the transmission probability is evaluated by breaking up the barrier into a series of discrete rectangular steps, and applying the exact solution method discussed in Section 24.2 for a series of steps. This general method to solve for transmission probabilities is called the **T-matrix** method. We will not present the method, but state and explain the result. The electron that tunnels into a state of energy E_i inside the barriers has a transmission amplitude T_1 to escape to the left, and T_2 to escape to the right. The lifetime of the electron inside the well is therefore $\tau = \frac{\hbar}{E_i(T_1+T_2)}$. In Chapter 29 we will see that the behavior of the electron here is mimicked by a photon inside a laser cavity. The net tunneling probability from the left to the right and the resulting net tunneling current are then given by

$$
\begin{aligned}
|T(E_z)|^2 &\approx \frac{4T_1T_2}{(T_1+T_2)^2}\cdot\frac{(\frac{\hbar}{\tau})^2}{(E_z-E_i)^2+(\frac{\hbar}{\tau})^2} \\
J_{RTD} = \frac{qg_sg_v}{L^d}\sum_k \Big[\frac{\hbar k_z}{m_c^\star}|T_k|^2 f_L(k)(1-f_R(k) &- \frac{\hbar k_z}{m_c^\star}|T_k|^2 f_R(k)(1-f_L(k)\Big] \\
&= \frac{qg_sg_vm_c^\star k_bT}{(2\pi)^2\hbar^3}\int_0^\infty dE_z\cdot|T(E_z)|^2\cdot\ln\Big[\frac{1+e^{\frac{F_L-E_z}{k_bT}}}{1+e^{\frac{F_L-E_z-qV}{k_bT}}}\Big],
\end{aligned}
\tag{24.59}
$$

where F_L is the quasi-Fermi level of the left side. Fig. 24.23 shows the experimentally measured tunneling current density as a function of the applied voltage for 300 K and 4 K, exhibiting a NDR in the forward-bias. The measured data proves that this genuine quantum behavior is robust and persists to room temperature, and has a weak temperature dependence. The corresponding model curve calculated using Equation 24.59 is in agreement with the experiment. The example RTD shown here is for a polar heterostructure. RTDs made from non-polar semiconductor heterostructures (such as AlGaAs/GaAs) exhibit symmetric NDRs at both forward- and reverse-bias conditions: the current-voltage curves are symmetric, quite unlike Esaki diodes. The NDR allows for the fabrication of high-speed RTD oscillations: frequencies of RTD oscillators can exceed 1 THz at room temperature when properly designed. Resonant tunneling is used in the design and operation of quantum cascade lasers, which is discussed in Chapter 29.

24.12 Bardeen's tunneling theory

Bardeen introduced an alternate method to capture electron tunneling transport from the point of view of transition rates. Though he introduced it to explain many-particle, correlated tunneling of Cooper pairs across superconducting junctions, the theory was adapted for single-particle tunneling by Harrison. This formulation is especially useful for modeling tunneling transport say between adjacent layers of 2D semiconductors that do not have a well-defined bandstructure connecting them. The main idea for tunneling from region a → region b is to evaluate the rate of electron scattering using Fermi's golden rule:

$$
\begin{aligned}
I &= q\times\Big[\frac{1}{\tau_{a\to b}}\cdot f_a\cdot(1-f_b) - \frac{1}{\tau_{a\to b}}\cdot f_b\cdot(1-f_a)\Big] \\
\frac{1}{\tau_{a\to b}} = \frac{2\pi}{\hbar}|\langle b|W|a\rangle|^2\delta(E_a-E_b), &\text{ and } \frac{1}{\tau_{a\to b}} = \frac{2\pi}{\hbar}|\langle a|W|b\rangle|^2\delta(E_b-E_a) \\
\implies &\boxed{I = \frac{2\pi q}{\hbar}\times\sum_{a,b}|\langle b|W|a\rangle|^2\delta(E_a-E_b)(f_a-f_b),}
\end{aligned}
\tag{24.60}
$$

where the usual identities $|\langle b|W|a\rangle|^2 = |\langle a|W|b\rangle|^2$ and $\delta(E_a - E_b) = \delta(E_b - E_a)$ are used. Furthermore, the density of states of the two sides $g_a(E) = \sum_a \delta(E - E_a)$ and $g_b(E) = \sum_b \delta(E - E_b)$ are then used to write the tunneling current as

$$I = \frac{2\pi q}{\hbar} \int dE |W(E)|^2 (f_a - f_b) g_a(E) g_b(E), \text{ where}$$
$$W(E) = \frac{\hbar^2}{2m_T^\star} \int_{\mathbf{S}} d\mathbf{S} \cdot (\psi_a^\star \nabla_r \psi_b - \psi_b^\star \nabla_r \psi_a). \tag{24.61}$$

Here $m_T^\star$ is the tunneling effective mass, the integral is evaluated over any surface **S** of constant potential energy inside the tunneling barrier, and ψ_a, ψ_b are the wavefunctions on the two sides of the tunneling barrier assuming the barrier extends infinitely to the other side. In the limit of a thick barrier, $W(E)$ becomes proportional to the WKB tunneling probability. In this form, the tunneling current is in a particularly expressive form with the product of the two density of states, and an integral over all possible energy states. The Fermi difference function $f_a - f_b$ defines the energy window within which tunneling occurs. If the density of states of one side (say $g_a(E)$) is known, the measurement of the tunneling current provides a method to measure the density of states $g_b(E)$ of the other side. This is the principle of a rich range of **tunneling spectroscopies**, of which the method of scanning tunneling microscopy (STM) is a prime example.

24.13 Kubo formalism

The theoretical models of electron tunneling treated up till now in this chapter have assigned particular shapes (e.g. triangular) to the potential barrier seen by electrons in the tunneling process. The net current and tunneling conductivity depends on the WKB tunneling probability through this barrier. This barrier shape is reasonable for metal-semiconductor junctions, or for intraband tunneling (e.g. in RTDs). For interband tunneling, they are plausible but admittedly arbitrary. The arbitrary assignment of a potential barrier for interband tunneling can be avoided by adopting the Kubo formalism, and evaluating the macroscopic response of the semiconductor to an external electric field[10]. In Chapter 20, Section 20.4, it was mentioned that the Kubo formalism used for DC electrical conductivity can be applied to explain tunneling conductivity. In fact the Kubo formalism does more: it shows that the DC electrical conductivity due to interband tunneling transport is intimately related to the modification of the optical absorption coefficient of the semiconductor in the presence of a large electric field, or the Franz–Keldysh effect, which is discussed in Chapter 27, Section 27.11.

We briefly outline this process here and suggest further reading materials at the end of the chapter. Writing the electric field as $F(t) =$

[10] This approach was outlined in R. Enderlein and K. Peuker, Physica Status Solidi (b), vol. 48, p. 231 (1971).

$F(\omega)e^{-i\omega t}$, the resulting electrical current flowing in the semiconductor $J(t)$ in the linear response approximation has a Fourier coefficient $J(\omega) = \tilde{\sigma}(\omega, F)F(\omega)$. The Kubo formula for the complex conductivity $\tilde{\sigma}(\omega, F)$ is expressed in terms of the correlation function:

$$\tilde{\sigma}(\omega, F) = i\frac{q^2 N}{m_e(\omega + i\gamma)}\hat{1} + \frac{1}{\hbar(\omega + i\gamma)}\int_0^\infty dt e^{i\omega t}\langle[j(0), j(-t)]_-\rangle_\infty,$$

$$j(t) = e^{\frac{iHt}{\hbar}} j(0) e^{\frac{-iHt}{\hbar}}, \text{ with } H = \sum_{i=1}^{N}[\frac{\mathbf{p}_i^2}{2m_e} + V(\mathbf{r}_i)] + H_{int} + q\sum_{i=1}^{N}\mathbf{F}\cdot\mathbf{r}_i. \tag{24.62}$$

In the above N is the number of electrons and γ is a scattering rate, making the first term provide the intraband conductivity for $\omega = 0$. The second term with the correlation function yields the interband tunneling current: for its derivation involving the density matrix see the references mentioned in ”Further reading” at the end of the chapter. The final result of the tunneling current $J_T(\omega = 0) = \int_0^F dF'\tilde{\sigma}(0, F')$ is

$$J_T(\omega = 0) = \frac{6p_{cv}^2}{m_e E_g}\frac{q(m_r^\star)^2 u^2 E_g^2}{\hbar^3 m_e}[\mathrm{Ai}'^2(u) - u\mathrm{Ai}^2(u)] \times (f_v - f_c), \tag{24.63}$$

where $u = \hbar\Theta/E_g$, and $\Theta = (\frac{q^2F^2}{2m_r^\star\hbar})^{\frac{1}{3}}$ is a characteristic frequency and p_{cv} is the familiar interband momentum matrix element from the $\mathbf{k}\cdot\mathbf{p}$ theory of Chapter 12. The final analytical solution involving Airy functions can be shown to be similar to those obtained earlier in this chapter using simpler approaches. The difference from earlier methods of this chapter is not having to choose a potential barrier. It is useful to appreciate the similarity of the final electrical tunneling conductivity to the optical electro-absorption coefficient as captured in the Franz–Keldysh effect discussed in Chapter 27, Section 27.11.

24.14 Landau–Zener theory

An early formulation of the dynamics of tunneling was by Zener (Fig. 24.14) and independently by Landau (Fig. 25.4). Their approach[11] is commonly referred to as **Landau–Zener** tunneling model, in which the tunneling problem is mapped to the solution of the time-dependent Schrödinger equation of a two-level quantum system. In this model, the process of tunneling is viewed as a transition from initial state $|0\rangle$ to the final state $|1\rangle$ as indicated in Fig. 24.24 (a). At time t the energies are $E_0(t)$ and $E_1(t)$, which depend explicitly on time via an external perturbation that tries to cross the levels as shown by dashed lines. Due to level repulsion, the crossing is avoided. The quantum state at time t is the superposition $a_0(1)|0(t)\rangle + a_1(t)|1(t)\rangle$. If at time $t = -\infty$ an electron is in state $|0\rangle$ and it stays *the same state* at $t \to \infty$, it is said

[11] This result was obtained also nearly at the same time by Ettore Majorana and Ernst Stueckelberg.

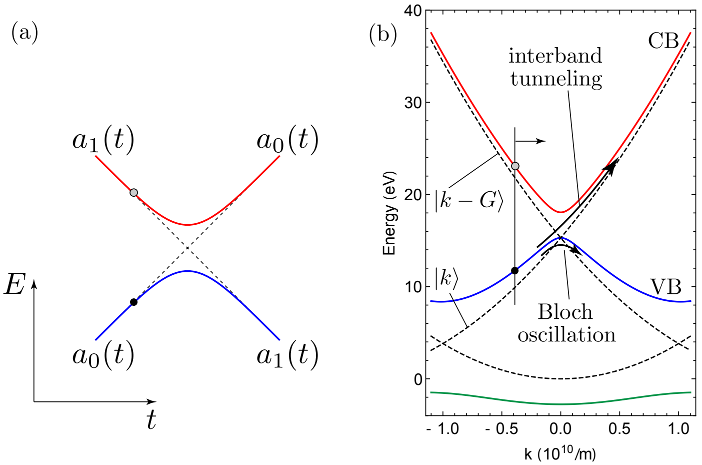


Fig. 24.24 The Landau–Zener approach to interband tunneling as a time-dependent perturbation problem.

to have tunneled across the gap of avoided crossing. In other words, the tunneling probability is $T = |a_0(t \to \infty)|^2$.

To apply the Landau–Zener model to semiconductors, consider the nearly free-electron (NFE) model (with no crystal potential!) of the semiconductor bandstructure as the exactly solved problem dashed lines in Fig. 24.24 (b). Consider the two bands labeled $|k\rangle$ and $|k-G\rangle$ where $G = 2\pi/a$ is a reciprocal lattice vector characterizing the periodic crystal potential of lattice constant a, which is $V(r) = \sum_G V_G e^{iGr}$. The periodic crystal potential splits the bands into the conduction and valence bands labeled CB and VB. For a electron starting out in the valence band state shown by the black dot, in the absence of scattering, the force due to the electric field $F = \hbar dk/dt$ sweeps the wavevector linearly with time as $k(t) = k(0) + \frac{F}{\hbar}t$. Both of these perturbations to the nearly free–electron bandstructure may be treated on equal footing in the following way.

The state of the system at time t is the superposition $|\psi(t)\rangle = a_0(t)|k(t)\rangle + a_1(t)|k(t)-G\rangle$ between the $|k(t)\rangle$ and $|k(t)-G\rangle$ states. Note that state $|k(t)\rangle$ is in the VB for $k < 0$ but in the CB for $k > 0$. The electric field sweeps the k-state linearly with t as shown, with the coefficients $a_0(t)$ and $a_1(t)$ changing with time. If upon crossing $k = 0$ the electron changes from NFE band $|k\rangle$ to NFE band $|k-G\rangle$, it stays in the valence band of the semiconductor and executes Bloch oscillations. But if it stays in the NFE band $|k\rangle$, it undergoes interband tunneling to the CB. The probability of this tunneling is obtained by solving the time-dependent Schrödinger equation written in the matrix form for the coefficients $a_0(t)$ and $a_1(t)$:

$$i\hbar\frac{d}{dt}\begin{bmatrix} a_0(t) \\ a_1(t) \end{bmatrix} = \begin{bmatrix} \frac{\hbar^2 k(t)^2}{2m_e} & -V_G \\ -V_G & \frac{\hbar^2 (k(t)-G)^2}{2m_e} \end{bmatrix}\begin{bmatrix} a_0(t) \\ a_1(t) \end{bmatrix}, \tag{24.64}$$

where $k(t) = k(0) - \frac{qF}{\hbar}t$ and $G = 2\pi/a_0$ is the reciprocal lattice constant. This coupled differential equation is deceptively difficult to solve analytically. Numerically, it exhibits oscillatory behaviour in time representing tunneling attempts. An asymptotic analytical solution is possible however for $|a_0(t \to +\infty)|^2$, which is precisely the interband tunneling probability. It is the famous Landau–Zener formula $T = |a_0(+\infty)|^2 = e^{-2\pi\Gamma}$, where $\Gamma = \frac{V_g^2}{\hbar\alpha}$ where $\alpha = \frac{\partial(E_1(t)-E_0(t))}{\partial t} = \frac{\hbar G q F}{m_e}$. In the two-band model of the semiconductor, using the effective mass relation $\frac{m_e}{\hbar G} = \sqrt{\frac{m_r^\star}{2E_g}}$ where $m_r^\star$ is the reduced effective mass and E_g the energy bandgap, the Landau–Zener interband tunneling probability becomes $T = \exp\,[-\frac{m_e a E_g^2}{4\hbar^2 qF}] = \exp\,[-\frac{\pi}{2\sqrt{2}}\frac{\sqrt{m_r^\star}E_g^{3/2}}{\hbar qF}] = \exp\,[-\frac{F_0}{F}]$. This is identical to the WKB forms obtained earlier, except for the numerical factor $\frac{\pi}{2\sqrt{2}} = 1.11$ instead of $\frac{4\sqrt{2}}{3} = 1.89$ in the exponent that was obtained for a triangular barrier model in Equation 24.47. If a parabolic barrier is assumed in WKB, the same prefactor in the exponent as the Landau–Zener result is obtained.

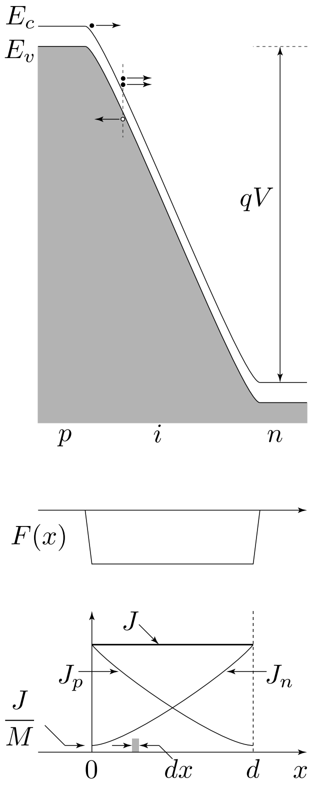


Fig. 24.25 Impact ionization in a semiconductor p-i-n diode. This process is also called **pair production** in high-energy physics jargon. A comprehensive review on the physics of impact ionization is Federico Capasso, "Physics of avalanche photodiodes." Semiconductors and Semimetals 22 (1985): 1-172. Derivation of the ionization integral for different electron and hole ionization coefficients is given in Stillman, G. E., and C. M. Wolfe. "Avalanche photodiodes." Semiconductors and Semimetals, Vol. 12. Elsevier, 1977. 291-393.

24.15 Avalanche processes and impact ionization

As shown in Fig. 24.19, for p-i-n diodes of moderate doping, **impact ionization** causes **avalanche breakdown** to occur before Zener breakdown. This is the breakdown mechanism in most practical diodes and transistors used in power electronics. This process is illustrated in the energy band diagram shown in Fig. 24.25 at a large reverse-bias. Minority electrons in the conduction band of the p-side on the left incident on the high-field *i*-region are accelerated to high kinetic energies due to the very high field. Above a critical value of kinetic energy, this **primary** electron can collide with an electron in the valence band and excite (or ionize) it into the conduction band upon impact. The result of the impact ionization is therefore an extra electron-hole pair, which are called the **secondaries**, as shown in the plane indicated by a dashed line in the energy band diagram.

Successive repetition of this process results in carrier multiplication: the number of electrons reaching the *n*-side is M times that enter from the *p*-side due to this avalanche process. Here M is the carrier multiplication factor. When $M \to \infty$, the diode reaches avalanche breakdown. We outline a procedure to quantitatively evaluate M and the breakdown process. Consider the spatial profile of the current density due to electrons and holes shown in Fig. 24.25. The total current $J = J_n + J_p$ at all x. At the *n*-electrode at $x = d$, the current is nearly entirely due to electrons: $J = J_n(d)$. If the multiplication factor is M, the electron current at $x = 0$ is nearly equal to $J_n(0) = J/M$. The multiplication of carriers is modeled as a generation rate $qG_n = \alpha_n J_n + \alpha_p J_p$ where $\alpha_n = \frac{1}{n}\frac{dn}{dx}$ and $\alpha_p = \frac{1}{p}\frac{dp}{dx}$ are the electron and hole impact ionization (or gain) coefficients in cm^{-1} units. By the continuity equation

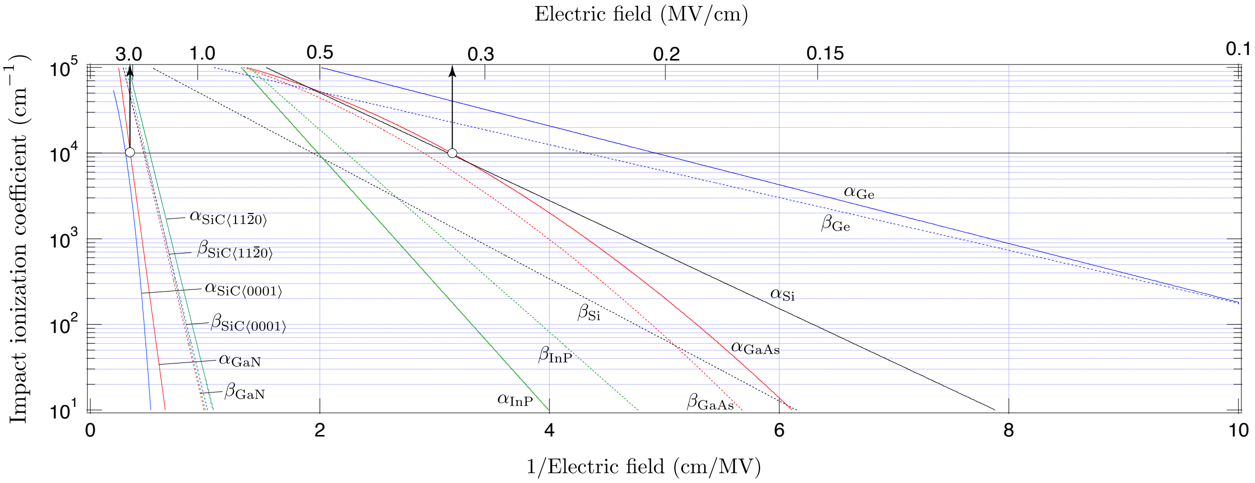


Fig. 24.27 Impact ionization coefficients of electrons ($\alpha = \alpha_n = \alpha_e$), and holes ($\beta = \alpha_p = \alpha_h$) of semiconductors as a function of the inverse electric field. Data and corresponding coefficients in Table 24.1 are kindly shared by Dr. Takuya Maeda, Cornell University.

$\frac{\partial n}{\partial t} + \frac{1}{q}\frac{dJ_n}{dx} = G_n$, in steady state we have

$$\frac{dJ_n}{dx} = \alpha_n J_n + \alpha_p J_p \underbrace{=}_{\text{for } \alpha_n=\alpha_p=\alpha} \alpha J \implies \frac{1}{J}\int_0^d dJ_n = \boxed{1 - \frac{1}{M} = \int_0^d \alpha dx.} \tag{24.65}$$

Though in general the impact ionization coefficients $\alpha_n \neq \alpha_p$, we first illustrate the physics of impact ionization assuming them to be the same. The boxed form of Equation 24.65 gives an expression for the multiplication factor M. Breakdown occurs for $M \to \infty$. This happens when $\int_0^d \alpha dx = 1$, which is the criterion for avalanche breakdown. In this integral, the impact ionization coefficient α depends on the distance x via the local electric field $F(x)$, and hence the dependence of α on F is the critical parameter that determines the carrier multiplication factor and the onset of avalanche breakdown. Fig. 24.27 shows the experimentally determined impact ionization coefficients α_n and α_p for various semiconductors. Their analytical forms are $\alpha = A \exp\left[-(B/F)^\delta\right]$, where A and B are constants and F is the electric field. The factor $\delta = 1$ for several semiconductors. The values of A, B, and δ for electron and hole impact ionization coefficients are listed in Table 24.1.

For a typical *i*-region or depletion region thickness of $d \approx 1\ \mu$m with a constant electric field, the breakdown criterion simplifies to $\alpha \approx 1/d \approx 10^4$ cm^{-1}, which gives a crude approximation for critical breakdown fields F_{cr} from Fig. 24.27. For example, for silicon it is $\gtrsim 0.3$ MV/cm and for GaN it is $\gtrsim 3.0$ MV/cm – these values should be only considered to be approximate based on the device structure.

Fig. 24.26 Leonid Keldysh made important contributions to the quantum mechanical understanding of interband tunneling and impact ionization processes in semiconductors. He formulated an important formalism for quantum field theory and many-body physics.

The impact ionization coefficients of semiconductors are evaluated directly from the **full bandstructure**. Typically the Monte Carlo technique is used with the full bandstructure, with the rates of phonon and defect scattering and that of pair production included stochastically. The rate of impact ionization (or pair-production) in semiconductors was evaluated by Keldysh (Fig. 24.26). Here we present a simple model that justifies why the impact ionization coefficients should of the form $\alpha = A \exp[-B/F]$.

Table 24.1 Impact ionization coefficients $\alpha = A \exp[-(B/F)^{\delta}]$. The coefficient A is in units of 10^6 cm^{-1}, and B is in 10^6 V/cm. Subscript e is for electrons and h is for holes.

A_e	B_e	δ_e	A_h	B_h	δ_h
Ge					
0.49	0.79	1	0.22	0.71	1
Si					
0.92	1.45	1	0.24	1.64	1
GaAs					
0.19	0.58	1.82	0.22	0.66	1.75
InP					
9.2	3.44	1	4.3	2.72	1
GaN ⟨0001⟩					
26.9	3.44	1	4.32	2.72	1
4H-SiC ⟨0001⟩					
0.14	4.93	1.02	3.14	11.8	2.37
4H-SiC ⟨11$\bar{2}$0⟩					
6.4	12.5	1	6.0	13.3	1

Consider the plane indicated by the dashed line in Fig. 24.25 where impact ionization is initiated by electrons. Assuming the valence band energy to be zero and parabolic bands, the incident "primary" electron has total energy $E_g + \frac{1}{2}m_c v_i^2$. After the impact ionization process, the primary electron energy is $E_g + \frac{1}{2}m_c^\star v_f^2$, the secondary conduction band electron has energy $E_g + \frac{1}{2}m_c^\star v_n^2$ and the valence band hole has energy $\frac{1}{2}m_v^\star v_p^2$. The impact ionization process conserves both momentum and energy. This requires, in a 1D approximation,

$$\begin{aligned} m_c^\star v_i &= m_c^\star v_f + m_c^\star v_n + m_v^\star v_p \\ \frac{1}{2}m_c^\star v_i^2 &= E_g + \frac{1}{2}m_c^\star v_f^2 + \frac{1}{2}m_c^\star v_n^2 + \frac{1}{2}m_v^\star v_p^2 \\ \text{for } v_n = v_p \implies (\frac{1}{2}m_c^\star v_i^2)_{min} &= E_{min} = \frac{2m_c^\star + m_v^\star}{m_c^\star + m_v^\star} E_g, \end{aligned} \tag{24.66}$$

where the threshold kinetic energy E_{min} of electrons to cause impact ionization is identified to be larger than the bandgap. For $m_c^\star = m_v^\star$, $E_{min} = \frac{3}{2}E_g$. The impact ionization coefficient α must be proportional to the number of electrons that have kinetic energy equal, or larger than this threshold energy. Assuming a Maxwell–Boltzmann distribution of carriers in the conduction band, the number of carriers of energy E_{min} or larger is proportional to $e^{-E_{min}/k_b T_e}$, where T_e is the hot-electron temperature.

The temperature T_e is obtained by the energy rate equation $\frac{dE}{dt} = qFv_d - \frac{k_b T_e}{\tau}$, where $k_b T_e$ is the electron energy, $1/\tau$ is the energy relaxation rate, and v_d is the electron velocity. In steady state, $k_b T_e = qFv_d\tau$, which makes the number of carriers that can cause impact ionization to be proportional to $e^{-E_{min}/qv_d\tau F}$, and as a result $\alpha \propto e^{-B/F}$, where $B = E_{min}/qv_d\tau$. The B coefficient is clearly in the units of an electric field, whose magnitude is of the order of the bandgap divided by a mean free path $v_d\tau$. This mean free path is typically that of optical phonon emission, and is typically $1 - 10$ nm in magnitude for most semiconductors. For semiconductors in which $v_d\tau$ depends on the electric field, the impact ionization coefficient goes as $\alpha = A\exp[-(B/F)^{\delta}]$, which is a general form as indicated in Table 24.1. The simplified discussion is put on more rigorous ground by a consideration of the bandstructure and scattering rates evaluated from Fermi's golden rule. Since the rate of phonon emission increases at higher temperature, the number of electrons that meet the impact ionization criteria is reduced. This leads to a **lowering** of the impact ionization coefficient with increase in temperature. The result is that the avalanche breakdown voltage is higher at higher temperature. For pn diodes that are doped heavily above $\sim 10^{19}/\text{cm}^3$, the depletion region thickness is too small to allow substantial carrier multiplication. In such diodes, Zener breakdown occurs before avalanche breakdown. For most diodes at lower doping densities the avalanche process is the dominant breakdown mechanism[12].

24.16 Tail of the dragon

Impact ionization does not need to be destructive: with a suitable current limit, the avalanche process can be controlled. The physical mechanism of the avalanche process in pn diodes has a close analogy to nuclear fission. In fission, a primary neutron collides with the nucleus of a heavy atom, and splits it into lighter nuclei, releasing a bit of energy $E = mc^2$, where m is the reduction of mass. This process also releases more secondary neutrons, leading to a neutron avalanche process. If the avalanche process is not restricted, the chain reaction leads to a nuclear explosion. If the chain reaction is controlled, the nuclear energy can be slowly extracted over elongated periods of time, forming the basis of reactors in nuclear power plants[13]. Controlled carrier multiplication by impact ionization is the basis for several semiconductor devices such as the avalanche photodiode (Chapter 28) and the **imp**act-ionization **a**valanche **t**ransit-**t**ime diode, or the IMPATT diode for generation of high-frequency power in the microwave domain.

24.17 Chapter summary section

In this chapter, we learned:

- Exact and the WKB approximation methods for tunneling probability of electrons in semiconductors through potential barriers,
- The nature of tunneling currents in metal-semiconductor Schottky diodes, and in pn junction diodes,
- The phenomena of interband tunneling and negative differential resistance in heavily doped pn junctions,
- NDR due to intraband resonant tunneling,
- Refined methods to capture the various nuances of tunneling including its time-dependent behavior, and
- Impact ionization phenomena and the resulting avalanche breakdown process in semiconductors.

---

[12] The empirical experimental dependence of the multiplication factor M on diode voltage V is $M = \frac{1}{1-(\frac{V}{V_{br}})^m}$ where m varies from $3 - 3.6$ for most semiconductors, and V_{br} is the breakdown voltage of the diode.

[13] Historically, the process of a controlled chain reaction was aptly named **"tickling the tail of the dragon"** by Feynman.

Further reading

Tunneling Phenomena in Solids edited by E. Burstein and S. Lundqvist is a comprehensive account of tunneling in semiconductors as well as in superconductors. **Tunneling in Solids** by C. B. Duke provides an exhaustive survey of tunneling theories including Bardeen's method, and a wealth of experimental phenomena. **Electronic Semiconductors** by E. Spenke is an early text that describes tunneling based on Bloch oscillations. **Quantum Mechanical Tunnelling and its Applications** by D. K. Roy describes the WKB and Bardeen methods and discusses several barrier shapes and nuances. **Physics of Low-Dimensional Semiconductors** by J. Davies has a detailed description of the T-matrix approach, and several details of resonant tunneling physics. **Advanced Physical Models for Silicon Device Simulation** by A. Schenk presents an excellent discussion of the Kubo formalism for interband tunneling, and several refined models for single and double barrier tunneling that appear in semiconductor devices.

Exercises

(24.1) **Tunneling probability in a metal-semiconductor junction**

In the derivation of the tunneling probability, it was mentioned that a prefactor appears in the calculation of the tunneling probability when calculated using solutions of the Schrödinger equation.

(a) The solution of the Schrödinger equation appears in the form of Airy functions. Read from the references mentioned at the end of the chapter and show how the prefactors appear as a result.

(b) Evaluate the more exact Fowler–Nordheim tunneling current density by retaining the prefactor and evaluate the Fowler–Nordheim tunneling current density numerically with and without the prefactor as given in Equation 24.45. Comment on how accurate it is to neglect the prefactor.

(c) As was discussed in Chapter 18, there is a rounding-off and lowering of the Schottky barrier due to image force. This modifies the expressions further: discuss the effect of image force lowering on the Fowler–Nordheim current. Sometimes this effect is accounted for by absorbing the change in an effective barrier height.

(24.2) **Fowler–Nordheim tunneling: approximations**

This question helps explain the derivation of the Fowler–Nordheim tunneling current density in Equation 24.46.

(a) Provide the missing steps of the derivation till the exact energy integral.

(b) Calculate and plot the supply function $N(E_z)$ as a function of E_z for $E_{Fm} = 10$ eV and for various $4\text{K} \leq T \leq 500$ K . How good is the $T \to 0$ K low-temperature approximation?

(c) Use the approximations to derive the boxed form of Equation 24.46. Show which terms have been neglected to obtain the boxed form, and justify why it is OK to neglect them.

(d) Make $\log(J/F^2)$ vs. $1/F$ plots of the numerically calculated exact temperature-dependent Fowler–Nordheim current and the boxed $T \to 0$ K current for $q\phi_b = 1.0$ eV, $m^\star = 0.2m_e$ for fields $0.01 \leq F \leq 10$ MV/cm in a log-log scale for $T = 4, 77, 300, 500$ K. Comment on how good is the simpler analytical $T \to 0$ K limit of the current for use at room temperature.

(24.3) **Vanishing act: tunneling escape in a flash memory**

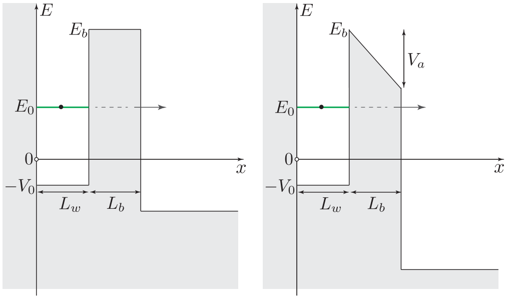


Fig. 24.28 Escape and field–emission by tunneling.

Fig. 24.28 shows a 1D potential for an electron, which is in the state with energy E_0 at $t = 0$. Since there is a lower potential for $x > L_w + L_b$, the state $|E_0\rangle$ is a *quasi-bound* state. The electron is destined to leak out.

(a) Using WKB tunneling probability, and combining semi-classical arguments, find an analytical formula that estimates the time it takes for the electron to leak out. Find a value of this lifetime for $L_b \sim 3$ nm, $L_w \sim 2$ nm, $V_0 \sim 1$ eV, $E_0 \sim 2$ eV, and $E_b \sim 5$ eV. How many years does it take?

(b) This feature is at the heart of *flash memory*, which you use in computers and cell phones. Find an analytical expression that describes how the lifetime changes if a voltage V_a is applied across the insulator. Estimate the new lifetime for $V_a \sim 2.8$ V. This is the *readout* of the memory.

(c) Now try solving the *same* problem using Fermi's golden rule, using the Bardeen approach outlined in the chapter. Model the problem carefully so that you can apply Fermi's golden rule. Discuss your approximations and their validity.

(24.4) **Esaki diodes and NDR**
Equation 24.57 gave the generalized interband tunneling current density in d-dimensions. Answer the following questions:

(a) Consider that you are required to design a conductor based on interband tunneling in a fixed cross-sectional area. The current flow must be by tunneling, and you can choose only one semiconductor, but you are free to choose whether you introduce a 3D tunnel junction, a 1D array of 2D tunnel junctions, or a 2D array of 1D tunneling junctions. Which choice will maximize the total current density?

(b) What are the practical concerns of implementing 2D or 1D tunnel junctions in quantum-confined structures?

(24.5) **Variations of tunneling current density**
Explore the effect of bandgap variation on the interband tunneling current density in Equation 24.57. For the cases where 2D or 1D tunnel junctions are formed by nanoscale structures of 3D semiconductors, explain the effect of interface and surface roughness on the tunneling current density.

(24.6) **Negative differential resistance**
Equation 24.58 provides the recipe for the calculation of NDR in Esaki diode structures. Fig. 24.21 showed three examples of diodes of different doping densities. Explore the effect of doping on the shape of the NDR. Specifically, discuss the doping design that will maximize the curvature at the peak of the NDR.

(24.7) **Backward diodes**
The history of pn junctions is interesting in the sense that the earliest pn diodes actually were more conductive in the *reverse-bias*! Only when the donor and acceptor doping concentrations could be controlled to low densities, the diffusion current behavior $J = J_0(e^{qV/k_bT} - 1)$ could be observed.

Do some research on this interesting history and write a short summary of your findings on the refinement of our understanding of pn junction diode currents over the years. Esaki's 1974 Nobel lecture (Reviews of Modern Physics, vol. 46, p. 237 (1974)) is a valuable resource of this aspect of the history of semiconductor physics.

Running Circles: Quantum Magnetotransport

25

25.1 Magnetotransport: a résumé 645
25.2 Hall effect 648
25.3 Magnetoresistance 656
25.4 Nernst and Ettingshausen effects 659
25.5 Cyclotron resonance 660
25.6 Faraday rotation 663
25.7 Atoms: Bohr magneton and spin 664
25.8 Landau levels in solids 668
25.9 Shubnikov de Haas effect 672
25.10 The quantum Hall effect 679
25.11 Quantum Hall effect theories 681
25.12 Hierarchy of Hall effects 692
25.13 Chapter summary section 695
Further reading 696
Exercises 697

The effect of externally applied magnetic fields on the electron energies in atoms, molecules, and metals and semiconductors have played a major role in the development of the laws of quantum mechanics and solid state physics. The electrical conductivity and optical properties of semiconductors in response to magnetic fields were the earliest probes of their current carriers and bandstructure. In addition to clarifying how semiconductors are fundamentally different from metals, magnetotransport and magneto-optical measurements have revealed a rich range of new quantum phenomena in semiconductors – many that were rather unexpected! In this chapter, we discuss:

- What are the semi-classical effects of magnetic fields on the properties of semiconductors?
- What are the quantum effects of a magnetic field on the properties of electrons in atoms?
- What are the quantum effects of magnetic fields on the physical properties of semiconductors?
- Can quantum transport theory explain Shubnikov de Haas oscillations and the quantum Hall effect observed in semiconductors?
- How can magnetic fields be mimicked in semiconductors through topological aspects in bandstructure?

25.1 Magnetotransport: a résumé

The force of attraction of a magnetic material has held a special fascination for people since its discovery in ancient times (it is often associated with personality traits!). Oersted's discovery in the 1820s that a current carrying wire could deflect a magnetic needle, followed by Faraday's discoveries of the reverse process established a deep connection between electricity and magnetism. The relationship between magnetic and electric fields and light was established by Maxwell's equations (see Chapter 26). While Faraday's findings, and Maxwell's equations provided a method to produce magnetic fields from electric currents and build electromagnets and electric motors, how exactly a material could be *intrinsically* magnetic *without a current* remained a deep mystery till early 1900s, till after the discovery of the electron.

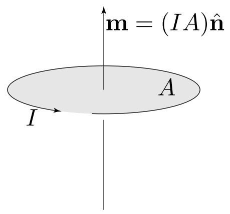


Fig. 25.1 In classical electromagnetism, a wire loop carrying a current I and enclosing an area A acts as a magnet of magnetic moment $\mathbf{m} = IA\hat{n}$.

The response of a material to an external magnetic field B is to itself become magnetic in one of many ways. The magnetization M of a solid of volume V is $M = -\frac{1}{V}\frac{\partial F}{\partial B}$ where $F = U - TS$ is the free energy at temperature T, U is the total internal energy, and S is the entropy. The magnetic susceptibility is $\chi_m = \frac{\partial M}{\partial B}$. Depending on the response of the material, it is said to exhibit paramagnetism ($\chi_m > 0$), diamagnetism ($\chi_m < 0$), ferromagnetism or ferrimagnetism ($\chi_m >> +1$), or antiferromagnetism ($\chi_m \approx 0$). Like many physical properties discussed in the earlier chapters, the first hints to understanding the origins of magnetism were provided by the discovery of the electron.

Fig. 25.2 Hendrika van Leeuwen proved that it is impossible for a material to exhibit magnetism if the particles in it follow the laws of classical physics. Magnetic properties of matter therefore are intrinsically quantum mechanical. The **Bohr–van Leeuwen theorem**, first anticipated by Bohr, is considered to be the driving force why Bohr broke with tradition to propose his orbit theory for the hydrogen atom that started quantum mechanics.

Langevin, who had studied in J. J. Thompson's laboratory where the electron was discovered, imagined that in atoms electrons moved in microscopic circular orbits as shown in Fig. 25.1, producing a magnetic moment $\mathbf{m} = IA\hat{\mathbf{n}}$, but with the vectors pointing in random directions at equilibrium. Since the energy of a magnet of moment $\mathbf{m}$ in an external magnetic field $\mathbf{B}$ is $-\mathbf{m}\cdot\mathbf{B} = -mB\cos\theta$, the net magnetization of N loops using a classical Maxwell–Boltzmann distribution is:

$$M = N \cdot \frac{\int_0^{\pi}(-m\cos\theta)e^{-\frac{mB\cos\theta}{k_bT}}\sin\theta d\theta}{\int_0^{\pi} e^{-\frac{mB\cos\theta}{k_bT}}\sin\theta d\theta} = N \cdot m \cdot \underbrace{(\coth x - \frac{1}{x})}_{\mathcal{L}(x)}, \quad (25.1)$$

where $\mathcal{L}(x)$ is the Langevin function and $x = (mB)/(k_bT)$. For $mB << k_bT$, $x << 1$, $\mathcal{L}(x) \approx \frac{x}{3}$, and $M \approx \frac{1}{3}\frac{Nm^2}{k_bT}B$. This yields a magnetic susceptibility $\chi_m = \frac{\partial M}{\partial B} \approx \frac{Nm^2}{3k_bT}$. With this model for **paramagnetism**, Langevin explained the experimental observation of his advisor Pierre Curie that the paramagnetism of oxygen gas increased at low temperatures. He fortuitously assumed all microscopic magnets to have the same magnetization (quantum mechanics later proved why this must be so). To explain **ferromagnetism** of solids, Weiss introduced the concept of an internal molecular field that added to the external magnetic field, making the total field $B \rightarrow B + \eta M$ proportional to the magnetization M. Building on Langevin's model, $M = Nm\mathcal{L}(\frac{m(B+\eta M)}{k_bT}) \approx \frac{Nm^2}{3k_bT}(B + \eta M)$, Weiss obtained the magnetic susceptibility $\chi_m = \frac{Nm^2}{3k_b(T-T_c)}$ which implies that the material develops a spontaneous magnetization for $T < T_c$, where $T_c = \frac{N\eta m^2}{3k_b}$ is the Curie temperature of the ferromagnetic material.

Fig. 25.3 Pieter Zeeman discovered that the sharp spectral lines of atoms were split by a magnetic field. His discovery was immediately recognized by Lorentz (Fig. 26.6) as proof that the atom housed a lighter particle that is more than a thousand times lighter than the hydrogen atom itself. This particle was shortly thereafter discovered to be the electron. Surprisingly, there is an entirely classical explanation for the Zeeman effect (with no $\hbar$!). Zeeman and Lorentz were together awarded the Nobel Prize in 1902 for this discovery.

The Langevin and Weiss models are for magnetism what the Drude model was for electrical conductivity. Similar to the Drude model, they have inconsistencies because of their dependence on classical mechanics and statistics. The most glaring inconsistency was pointed out in the **Bohr–van Leeuwen theorem**, which states that no magnetism can result from particles that follow classical mechanics and classical statistics (see Fig. 25.2). We prove this theorem later in this Chapter. **Magnetism is inherently a quantum mechanical property**.

In late 1890s, Pieter Zeeman (Fig. 25.3) discovered that a magnetic field split the characteristic sharp spectral lines of atoms into several sharp lines with separation proportional to the magnetic field. Bohr considered this **Zeeman effect** and the impossibility of magnetism in classical mechanics to introduce, in his orbit model[1], the concept of an elementary magnet. Analogous to an elementary negative charge (electron) or positive charge (proton), the quantum of magnetic moment of an electron is called the **Bohr magneton**, equal to $\mu_B = \frac{q\hbar}{2m_e}$.

Experiments performed by Stern and Gerlach in 1920s confirmed Bohr's predictions, but also uncovered an additional bit of magnetism (or angular momentum) that could not be associated with the orbital motion of electrons. This additional magnetic moment is an inherent property of the electron, inseparable from it just as its charge and its mass: it is the **spin** of the electron[2]. Therefore the conservation of total angular momentum must include both the orbital part and the spin part. The spin and orbital angular momenta are coupled and lead to **spin-orbit interactions** in various solids including semiconductors. Such interactions are relativistic in origin: in fact, the existence of electron spin was shown by Dirac to be the natural consequence of requiring quantum mechanics be consistent with special relativity.

In 1927, Pauli showed that the spin of electrons in metals, and the Fermi–Dirac statistics was responsible for the weak paramagnetism. The **Pauli quantum theory of paramagnetism** was the first application of Fermi–Dirac statistics to solid state, and is the quantum version of Langevin's classical theory. In 1928, Heisenberg gave the Weiss molecular field model of ferromagnetism a quantum facelift by introducing an **exchange interaction** between neighboring electron spins.

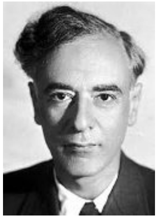


Fig. 25.4 Lev Landau, one of the most influential physicists of the 20th century. Among his contributions are the quantum theory of diamagnetism, the theories of phase transitions, and the Landau-Ginzburg theory of superfluidity and superconductivity. He was awarded the 1962 Nobel Prize in Physics. His masterful Landau–Lifshitz series of physics textbooks is an enduring legacy.

In 1930, a remarkable transport phenomena was discovered by de Haas and Shubnikov. They observed that the electrical resistivity of bismuth films showed **Shubnikov de Haas oscillations** as a function of magnetic field at high field and low temperature. Analogous **de Haas van Alphen oscillations** were also observed in the magnetic susceptibility χ_m of metals. In 1930, Landau (Fig. 25.4) solved the Schrödinger equation for free–electrons in a magnetic field, and showed that the density of states collapse into **Landau Levels**, which are the root of the quantum oscillations. The de Haas van Alphen effect was explained by Peierls, whereas the theoretical explanation of Shubnikov de Haas oscillations had to await till the late 1950s. This is because the quantum magnetotransport effects lay outside the scope of Boltzmann transport, and require a density-matrix approach. Quantum oscillations of magnetotransport and magnetic susceptibilities, and **cyclotron resonance** are valuable tools to measure bandstructures, effective masses, and Fermi surfaces of metals, semimetals, and semiconductors. **Magneto-optical spectroscopy** allows a method of measuring Landau level density of states in semiconductors. **Magnetophonon oscillations** are also observed in several semiconductors where the phonon energy can cause transitions between Landau levels.

Fig. 25.5 Klaus von Klitzing discovered the integer quantum Hall effect in 1980 with G. Dorda and M. Pepper. He was awarded the Physics Nobel Prize in 1985.

The classical Hall effect directly yields the electron and hole densities and their mobilities in semiconductors. Remarkable quantum effects were observed by von Klitzing (Fig. 25.5) in 1980. The **integer**

---

[1] As discussed in Chapter 3, the Bohr model quantizes the angular momentum of electrons, and consequently, also the magnetic moment of the atom. The Bohr orbit model failed to explain the effect of magnetic field on the spectra of atoms other than hydrogen. This was later resolved by quantum mechanics.

[2] This was first realized by Goudsmidt and Uhlenbeck in 1925. Spin was quantitatively understood soon thereafter by Pauli who introduced his eponymous spin matrices. The projection of the spin angular momentum along the direction of an external magnetic field can take only two values, of $+\frac{\hbar}{2}$ and $-\frac{\hbar}{2}$.

quantum Hall effect observed in semiconductor 2D electron or hole gases produces a Hall conductance that is so precisely quantized – up to parts per billion – that it forms the basis of the new fundamental resistance standard. Theories of gauge invariance of Laughlin, of edge modes by Buttiker, and of topological Chern number quantization of Thouless, Kohtomo, Nightingale, and den Nijs (TKNN) based on the Kubo theory were presented to explain the integer quantum Hall effect and the robustness of the quantization. Shortly thereafter, the **fractional quantum Hall effect** was discovered in high mobility semiconductor 2D electron gases, with plateaus at fractions with odd integer denominators. A theory of **composite fermions** has been developed to explain this phenomenon.

The quantized Hall effects introduced topology into condensed matter electronic and transport properties, in the form of uncovering possibilities of topologically protected transport. A hierarchy of Hall effects, such as the **spin Hall effect** and **quantum spin Hall effect** have been discovered. Magnetic effects of semiconductors form the basis of a large suite of practical characterization methods. In addition, they are used in Hall-sensors, and in magnetic memory. Magnetic effects are playing an increasingly important role in the design of qubits.

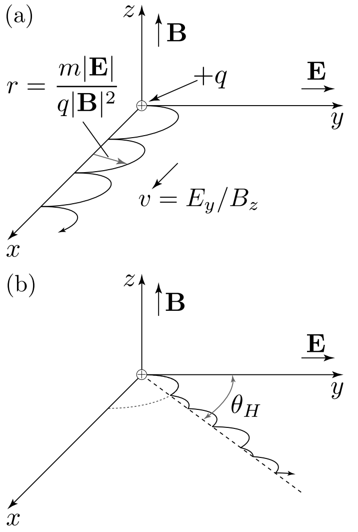


Fig. 25.6 Motion of a positive charge q and mass m in a crossed electric and magnetic field starting from rest at the origin in "infinitely large" samples: (a) motion in vacuum, and (b) motion impeded by scattering. In vacuum the motion is along the direction $\mathbf{E} \times \mathbf{B}$ which is perpendicular to the electric field. Scattering changes the "net" direction of current flow, which makes a **Hall angle** θ_H with the electric field, where $\tan\theta_H = \mu|\mathbf{B}|$. Here $\mu = q\tau/m$ is the Hall effect mobility and τ is an averaged scattering time.

25.2 Hall effect

Small magnetic fields cause perturbative effects to the transport of electrons and holes, resulting in the "classical" Hall effects and magnetoresistance. Large magnetic fields change the density of states of solids by quantizing the electron orbits of Bloch states, and therefore fundamentally alter the electronic and photonic properties of semiconductors. Before delving into the quantum effects of magnetic fields, we devote several sections to the discussion of semi-classical magnetotransport phenomena in semiconductors which were introduced in Chapter 21, but with a direct connection to experiments.

A completely classical picture of the effect of a magnetic field on charge transport in semiconductors is a powerful and intuitive guide to the more refined semi-classical Boltzmann-transport and the fully quantum-mechanical approaches that are discussed in later parts of this chapter. Consider a particle of charge $+q$ and mass m at rest at the origin as shown in Fig. 25.6 (a). An electric field $\mathbf{E} = E_y\hat{y}$ and a magnetic field $\mathbf{B} = B_z\hat{z}$ are turned on simultaneously. The electric field causes a force $\mathbf{F}_E = q\mathbf{E}$ on the particle. The moment the charge begins accelerating in the direction of the electric field with instantaneous velocity $\mathbf{v}(t)$, it experiences a force due to the magnetic field $\mathbf{F}_B = q\mathbf{v}(t) \times \mathbf{B}$ that pushes it towards the x-axis. The particle traces a curved trajectory which is completely determined by Newton's law $m\ddot{\mathbf{r}} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, written out in vector form:

$$m \begin{bmatrix} \dot{v}_x \\ \dot{v}_y \\ \dot{v}_z \end{bmatrix} = q(\begin{bmatrix} 0 \\ E_y \\ 0 \end{bmatrix} + \text{Det} \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ v_x & v_y & v_z \\ 0 & 0 & B_z \end{bmatrix}) \implies \begin{bmatrix} \dot{v}_x \\ \dot{v}_y \\ \dot{v}_z \end{bmatrix} = \underbrace{\frac{qB_z}{m}}_{\omega_c} \begin{bmatrix} v_y \\ \frac{E_y}{B_z} - v_x \\ 0 \end{bmatrix}, \tag{25.2}$$

where a characteristic **cyclotron frequency** $\omega_c = qB_z/m$ is identified. Since $\dot{v}_z = 0$, there is no change in the velocity along the z-direction along the **B** field. If the particle had an initial nonzero v_z, this component would not change with time. Since in Fig. 25.6 the particle starts from rest at the origin, its motion is confined in the $x - y$ plane. Since $\dot{v}_x = +\omega_c v_y$, we get $\ddot{v}_y = -\omega_c^2 v_y$. Then, the solution with initial conditions $v_y(0) = y(0) = v_x(0) = x(0) = 0$ is: $y(t) = \frac{E_y}{\omega_c B_z}(1 - \cos \omega_c t)$ and $x(t) = \frac{E_y}{\omega_c B_z}(\omega_c t - \sin \omega_c t)$. The resulting path traced is shown in Fig. 25.6 (a): it is a **cycloid**. For zero initial velocity the cycloid is a series of sections of circles[3], each of radius $r = E_y/(\omega_c B)$. On top of the oscillatory components that integrate in every time period to zero, an important feature of the motion of the particle in the crossed electric and magnetic field is a **non-oscillatory drift** in the $\mathbf{E} \times \mathbf{B}$ direction. Here it is the x-direction with a velocity $\langle \dot{x}(t) \rangle = v_H = E_y/B_z$: the net result is a **current flow** in the $+x$, or the $\mathbf{E} \times \mathbf{B}$ direction. For example, if $E_y = 1$ kV/cm and $B_z = 1$ T, the x-directed velocity is $v_H = E_y/B_z = 10^7$ cm/s, a high velocity by semiconductor standards[4].

However, all this is true only if three conditions are satisfied. First, the particle should not undergo any scattering. Second, it should not meet any sample ”boundaries”. Third, we have only assumed a single particle. In most semiconductor structures of practical interest, all three conditions can be violated. We will approach the problem in successive steps. Let us first investigate how the trajectory changes with scattering. Then, we investigate the changes with finite boundaries of the sample, assuming all particles move with the same velocity. In a real semiconductor there are a large number of particles whose initial velocities range over magnitudes and directions. To account for the distribution of carrier velocities, by employing the services of the Boltzmann transport equation, we will find that the simpler pictures explain several magnetotransport properties, but not all.

Effect of scattering: Fig. 25.6 (b) shows the effect of scattering on the trajectory. Every scattering event resets the velocity, and the particle starts along a new cycloidal trajectory. The net result of the successive scattering events is that the drift path is deflected. The drift velocity along **E** is $\mathbf{v} = \mu \mathbf{E}$ where μ is the particle mobility. The magnetic force is $\mathbf{F}_B = q\mu(\mathbf{E} \times \mathbf{B})$, and the total force on the particle is $q\mathbf{E} + q\mu(\mathbf{E} \times \mathbf{B})$. The total force is no longer perpendicular to both **E** and **B**, but at an angle θ_H with **E** such that for $\mathbf{E} \perp \mathbf{B}$ as shown in Fig. 25.6 (b),

$$\tan \theta_H = \frac{q\mu|\mathbf{E} \times \mathbf{B}|}{q|\mathbf{E}|} = \mu B, \text{ where } \theta_H \text{ is the } \textbf{Hall angle}. \tag{25.3}$$

The current flows along the direction of net force, making an angle

[3] If the initial velocity was non-zero, the path remains a cycloid, but of different shapes that may self-intersect.

[4] Note that the velocity is *independent* of the mass of the particle; it depends only on the externally applied fields.

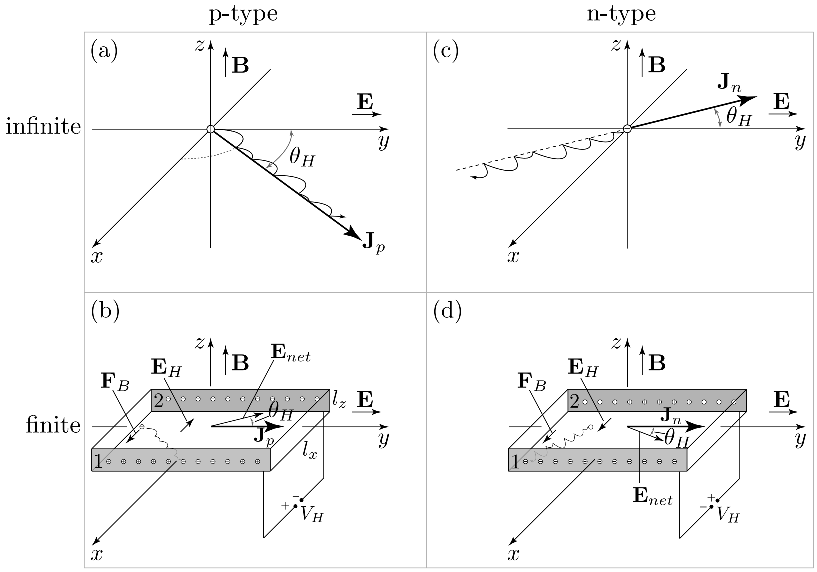


Fig. 25.7 Effect of magnetic field on (a, c) infinite, and (b, d) finite p-type (a, b) and n-type (c, d) semiconductors. (b, d) show the accumulation of charge at opposite faces and the development of a Hall voltage.

θ_H with **E**. Notice that for a given magnetic field B_0, a large mobility $\mu >> 1/B_0$ will make $\theta_H \rightarrow \frac{\pi}{2}$, approaching the case of no scattering. On the other extreme, for a low mobility $\mu << 1/B_0$, the Hall angle $\theta_H \rightarrow 0$ and the current flows almost along the electric field, nearly unaffected by the magnetic field[5].

Note that the units of mobility are that of inverse magnetic field, and μB is dimensionless. We can write $\mu B = \frac{q\tau}{m}B = \frac{qB}{m}\tau = \omega_c \tau$ which is the product of the cyclotron frequency and the scattering time, and the Hall angle is $\tan \theta_H = \omega_c \tau$. The meaning is transparent: if $\omega_c \tau << 1$, the scattering time is much smaller than one time period of the cycloid. The particle covers a small fraction of one cycloidal period before it is scattered again. On the other hand for $\omega_c \tau >> 1$, it can complete several periods of the cycloid before it scatters.

Thus, the effect of a magnetic field in the presence of scattering in an infinitely large sized sample is the rotation of the current density vector $\mathbf{J}_p$ with respect to the direction of the electric field **E** by the Hall angle θ_H as shown for holes in a p-type semiconductor in Fig. 25.7 (a). Fig. 25.7 (c) shows that the Hall angle between $\mathbf{J}_n$ and **E** for negatively charged particles is rotated in the opposite direction.

Effect of finite sample size: The effect of a finite sized sample is shown in Figures 25.7 (b) and (d). It is the **Hall effect**, arguably one of the most important experimental measurement techniques for understanding carrier transport in semiconductors. Consider Fig. 25.7 (b) for holes. Assume that the electric field **E** is turned on first without the magnetic field. The mobile holes of volume density p drift in the direction of the field. Let us assume for now that all p holes per unit volume move at the same velocity. Then, if the magnetic field **B** is turned on, the magnetic force $\mathbf{F}_B$ develops in the $+x$ direction,

[5] For example, at $B_0 = 1$ T, a mobility of 10^5 cm^2/V·s gives $\theta_H = 84.3°$ which is almost along the $\hat{x}$ axis, whereas a mobility of 10^2 cm^2/V·s gives $\theta_H = 0.57°$ which is almost along the y-axis in Fig. 25.6 (b)

which adds to the electric force $\mathbf{F}_E$ to push the mobile positive charges along the cycloidal trajectory shown towards the bottom face labeled "1". The initial process is similar to Fig. 25.7 (a), but once the positive charges reach face "1", they accumulate there. The opposite face labeled "2" is charged equal and negative to maintain charge neutrality. As the charges accumulate, an electric field $\mathbf{E}_H$ pointing from the $+$ve charges from face "1" to the $-$ve charges in face "2" builds up. For positive mobile charges, this process will continue till the net force in the x-direction $\mathbf{F}_B + q\mathbf{E}_H$ is exactly zero. When this happens, the system reaches steady state, and holes now experience no net force in the x-direction. The current vector $\mathbf{J}_p$ therefore points in the same direction as $\mathbf{E}$, but now the net electric field is $\mathbf{E} + \mathbf{E}_H = \mathbf{E}_{net}$, which points[6] at a Hall angle θ_H away from the direction of $\mathbf{J}_p$.

The importance of the Hall effect stems from the fact that the accumulation of charge is directly measurable as a Hall voltage V_H as indicated in Fig. 25.7 (b). This voltage is $V_H = E_H l_x$ where l_x is the distance between faces "1" and "2". We therefore get

$$\mathbf{F}_B + q\mathbf{E}_H = 0 = q\mathbf{v} \times \mathbf{B} + q\mathbf{E}_H = \frac{1}{p}\mathbf{J} \times \mathbf{B} + q\mathbf{E}_H \implies \mathbf{E}_H = \frac{1}{qp}\mathbf{B} \times \mathbf{J}$$

$$I = \mathbf{J} \cdot \mathbf{A} = J l_x l_z \implies R_H = \frac{1}{qp} = \frac{E_H}{BJ} \cdot \frac{l_x l_z}{l_x l_z} = \frac{V_H l_z}{IB} \implies \boxed{p l_z = \frac{IB}{qV_H}.} \tag{25.4}$$

In other words, the measurement of the Hall voltage V_H at a given current I and magnetic field B in the geometry of Fig. 25.7 (b) directly yields the product of the mobile hole density p and the thickness l_z, which is the integrated sheet density of mobile carriers. The quantity $R_H = \frac{1}{qp} = \frac{V_H l_z}{IB}$ is called the Hall coefficient[7] and is of units $\frac{\text{cm}^3}{\text{Coulomb}}$. Clearly, $R_H = \frac{\mu}{\sigma}$, which implies if the conductivity σ is known, the mobility μ is obtained from $\mu = R_H \sigma$. Thus by combining the Hall effect measurement with an electrical conductivity measurement, the carrier mobility is also obtained[8].

But the Hall effect measurement has one more ace up its sleeve. From Fig. 25.7 (d) it is already seen that the Hall voltage V_H is of opposite sign for electrons than for holes. Thus, for electrons of negative charge and volume density n, Equations 25.4 change to:

$$\mathbf{F}_B - q\mathbf{E}_H = 0 = \frac{1}{n}\mathbf{J} \times \mathbf{B} - q\mathbf{E}_H \implies \mathbf{E}_H = -\frac{1}{qn}\mathbf{B} \times \mathbf{J}$$

$$\implies R_H = -\frac{1}{qn} = \frac{V_H l_z}{IB} \implies \boxed{n l_z = -\frac{IB}{qV_H}.} \tag{25.5}$$

The measured Hall coefficient is then negative[9]. **Therefore, the Hall effect gives the density, mobility, *and* the sign of the mobile charge carriers in a semiconductor.** It is rather remarkable that the Edwin

---

[6] Thus, as opposed to the infinite sample where the magnetic field rotates the direction of the current density vector $\mathbf{J}_p$, for a finite sample a magnetic field rotates the net electric field $\mathbf{E}_{net}$ and not the current density vector.

[7] As an example, for $p = 10^{16}$ /cm^3, the Hall coefficient is $R_H = 625\frac{\text{cm}^3}{\text{Coulomb}}$. For a Hall conductor with carrier density $p = 10^{16}$ /cm^3 and of thickness $l_z = 1$ μm, at a injected current $I = 1$ mA and $B = 0.1$ T, the Hall voltage is $V_H = R_H \cdot \frac{IB}{l_z}$ = 62.5 mV. The actual measured voltage may be higher by a Hall factor of order unity due to the nature of the scattering mechanism, as discussed shortly.

[8] Drift and Hall mobilities will be differentiated shortly.

[9] Note that for holes the Hall voltage is $V_H = \frac{IB}{qpl_z}$, and for electrons $V_H = -\frac{IB}{qnl_z}$. Since I, B, l_z, q, n, p are all positive, the sign of the Hall voltage is the same as that of the mobile carriers.

Hall (Fig. 2.11) discovered this effect nearly *two decades before* the discovery of the electron itself! Even after the discovery of the electron and its negative charge, the positive charge carriers measured by Hall effect on several materials remained a thorn in the explanation of conductivity till the formulation of the band theory of solids and the understanding of holes as mobile carriers[10].

We have thus seen the effects of 1) scattering and 2) finite sample size on carrier transport in a magnetic field. Let us now remove the third restriction and allow for carrier distributions over several velocities and energies as is the case in semiconductors. In fact, since we will employ the Boltzmann transport equation, we can simultaneously enforce all conditions such as scattering as well as sample boundary conditions.

Effect of distribution of carrier velocities: Let us now recall what the Boltzmann transport equation (BTE) has to say about the effect of a "non-quantizing" magnetic field on charge transport in semiconductors. A vast array of fascinating effects result from the combination of electric fields, magnetic fields, and thermal gradients. These were introduced in Chapter 21, Equations 21.53 and 21.54, and a universal kinetic coefficient (or conductivity) to calculate the current density in response to one, or more of these driving forces was presented in Equation 21.55. These equations are discussed here again to understand several magnetotransport phenomena.

The effects of a magnetic field on charge and heat transport in the BTE are captured by an effective force vector **G** which was derived in Equation 21.52 derived in Chapter 21 with $\mathbf{F} = Q\mathbf{E}$ the electric force:

$$\mathbf{G} = \frac{1}{(1+\mu^2|\mathbf{B}|^2)}[\mathbf{F} + \mu(\mathbf{F}\times\mathbf{B}) + \mu^2(\mathbf{F}\cdot\mathbf{B})\mathbf{B}]$$

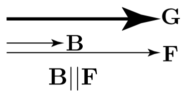
$$\mathbf{G} = |\mathbf{F}|\cdot[\underbrace{(\frac{1}{1+b^2})\hat{\mathbf{n}}_{\mathbf{F}}}_{\text{conductivity}} + \underbrace{(\frac{b}{1+b^2})\hat{\mathbf{n}}_{\mathbf{F}\times\mathbf{B}}}_{\text{Hall effect}} + \underbrace{(\frac{b^2}{1+b^2})\hat{\mathbf{n}}_{\mathbf{B}}}_{\text{magnetoresistance}}], \qquad (25.6)$$

where we recognize $b = \mu|\mathbf{B}| = \tan\theta_H$, where θ_H is the Hall angle. The components of the generalized force along the electric field and concentration gradients captured by **F** and the magnetic field **B** are made explicit by splitting it into three components: along **F**, **B**, and $\mathbf{F}\times\mathbf{B}$. Table 25.1 and the corresponding Fig. 25.8 show the orientation of **G** with respect to **F** and **B**. For a longitudinal magnetic field, there is no Hall effect force, though there may be a magnetoresistance. For a transverse magnetic field, there is a Hall effect force, but no magnetoresistance. The relative strengths of the three components of **G** are controlled by b, the product of the mobility and the magnetic field. Fig. 25.9 shows this dependence. The actual current depends on the **G** vector, the carrier distribution function $f^0_{\mathbf{k}}$ and the semiconductor bandstructure via the group velocity $\mathbf{v}_{\mathbf{k}}$ and the momentum scattering time $\tau_{\mathbf{k}}$ by the relation $\mathbf{j} = \frac{1}{L^d}\sum_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})\phi_{\mathbf{k}} =$

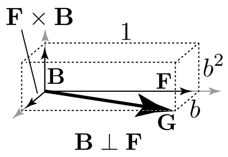
[10] The Hall effect highlights how careful experimental measurements reveal phenomena (e.g. of holes) that sometimes need a long time (here nearly half a century) to understand and explain. See Exercise 25.1 for further exploration of the origins of the Hall effect in a mistake made by Maxwell.

Table 25.1 **G** vector in a magnetic field.

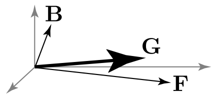
$\frac{\mathbf{G}}{\lvert\mathbf{F}\rvert} =$	$[(\frac{\hat{\mathbf{n}}_{\mathbf{F}}}{1+b^2})$	$+(\frac{b\hat{\mathbf{n}}_{\mathbf{B}\times\mathbf{F}}}{1+b^2})$	$+(\frac{b^2\hat{\mathbf{n}}_{\mathbf{B}}}{1+b^2})\ ]$
$\mathbf{B} = 0$	$[\hat{\mathbf{n}}_{\mathbf{F}}$	$+0$	$+0]$
$\mathbf{B}\lvert\rvert\mathbf{F}$	$[(\frac{\hat{\mathbf{n}}_{\mathbf{F}}}{1+b^2})$	$+0$	$+(\frac{b^2\hat{\mathbf{n}}_{\mathbf{F}}}{1+b^2})]$
$\mathbf{B}\perp\mathbf{F}$	$[(\frac{\hat{\mathbf{n}}_{\mathbf{F}}}{1+b^2})$	$+(\frac{b\hat{\mathbf{n}}_{\mathbf{B}\times\mathbf{F}}}{1+b^2})$	$+0]$



Longitudinal



Transverse



General

Fig. 25.8 Orientation of the **G** vector with respect to the electric vector **F** and the magnetic field **B**. The vector $\mathbf{F} = Q\mathbf{E} + \nabla_{\mathbf{r}}[E_F - E_c(\mathbf{r})]$ which consists both the electric field and the concentration gradients is purely $Q\mathbf{E}$ in a homogeneous semiconductor. If the magnetic field points along **F**, the b^2 terms cancel and $\mathbf{G} = \mathbf{F}$. For small perpendicular magnetic fields for which $b = \mu B << 1$, the **G** vector points mostly along the electric field vector **F**. The general case of large magnetic field and general orientation has a richer variation in magnitude and direction of **G**.

$\frac{1}{L^d}\sum_{\mathbf{k}} \mathbf{v}_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})(-\tau_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}\cdot\mathbf{G})$. The particle flux density for charge Q is

$$\mathbf{J}/Q = \mathbf{j} = [\underbrace{K_{10}\mathbf{F}}_{\propto\mathbf{E}} + \underbrace{K_{20}(\mathbf{F}\times\mathbf{B})}_{\text{Hall effect}} + \underbrace{K_{30}\mathbf{B}(\mathbf{B}\cdot\mathbf{F})}_{\text{magnetoresistance}}\ ]$$

$$+[\underbrace{K_{11}\frac{\nabla_{\mathbf{r}}T}{T}}_{\text{Seebeck}} + \underbrace{K_{21}(\frac{\nabla_{\mathbf{r}}T}{T}\times\mathbf{B})}_{\text{Nernst}} + \underbrace{K_{31}\mathbf{B}(\mathbf{B}\cdot\frac{\nabla_{\mathbf{r}}T}{T})}_{...}], \tag{25.7}$$

where $\mathbf{J}$ is the charge current density, and

$$\boxed{K^{ij}_{rs} = \frac{1}{L^d}\cdot(\frac{Q}{m^\star})^{r-1}\cdot\sum_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})\frac{v_i v_j \tau^r_{\mathbf{k}}(E_{\mathbf{k}}-\mu)^s}{1+(\frac{Q\tau_{\mathbf{k}}}{m^\star})^2|\mathbf{B}|^2}} \tag{25.8}$$

is the generalized kinetic coefficient K_{rs}. The corresponding generalized heat current $\mathbf{J}_q = \frac{1}{L^d}\sum_{\mathbf{k}}(E_{\mathbf{k}}-\mu)\mathbf{v}_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})(-\tau_{\mathbf{k}}\mathbf{v}_{\mathbf{k}}\cdot\mathbf{G})$ is

$$\mathbf{J}_q = [\underbrace{K_{11}\mathbf{F}}_{\text{Peltier}} + \underbrace{K_{21}(\mathbf{F}\times\mathbf{B})}_{\text{Ettingshausen}} + \underbrace{K_{31}\mathbf{B}(\mathbf{B}\cdot\mathbf{F})}_{...}]$$

$$+[\underbrace{K_{12}\frac{\nabla_{\mathbf{r}}T}{T}}_{-\kappa_e\nabla_{\mathbf{r}}T} + \underbrace{K_{22}(\frac{\nabla_{\mathbf{r}}T}{T}\times\mathbf{B})}_{\text{Righi–Leduc}} + \underbrace{K_{32}\mathbf{B}(\mathbf{B}\cdot\frac{\nabla_{\mathbf{r}}T}{T})}_{...}]. \tag{25.9}$$

With these relations, we shall evaluate all magnetoresistance coefficients. First, we clarify the combinations of $\mathbf{E}, \mathbf{B}, \nabla_{\mathbf{r}}T$ that contribute to the current density vectors $\mathbf{J}$ and $\mathbf{J}_q$. Why is there a term proportional to the electric force $\mathbf{F}$ and to $\nabla_{\mathbf{r}}T$ but not to the magnetic field $\mathbf{B}$ in Equations 25.7 and 25.9? In Fig. 25.10, it is argued why a magnetic field $\mathbf{B}$ is not a true vector: it is a pseudo-vector. In other words, $\mathbf{B}$ is an **axial vector**, as opposed to $\mathbf{E}$, $\nabla_{\mathbf{r}}T$, $\mathbf{J}$, and $\mathbf{J}_Q$ which are all **polar vectors**. They differ in their mirror reflection properties. This difference justifies why in Equations 21.53 and 21.54 the charge current density $\mathbf{J}$ and heat current density $\mathbf{J}_Q$ – which are polar vectors – were expressed in direct proportion to the polar vectors $\mathbf{E}$ or $\nabla_{\mathbf{r}}T$, but their dependence on the axial vector $\mathbf{B}$ is only through a combination of its scalar and vector products with $\mathbf{E}$ or $\nabla_{\mathbf{r}}T$.

To use the generalized kinetic coefficient for the calculation of the Hall effect, Equation 25.8 can be evaluated for general non-isotropic bandstructures. We limit the discussion here for a charge $Q = q$, and an isotropic bandstructure $v_i = v_j = v_{\mathbf{k}}$ and effective mass $m^\star$. The kinetic coefficients then become

$$K_{rs} = \frac{1}{L^d}(\frac{q}{m^\star})^{r-1}\sum_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})\frac{v^2_{\mathbf{k}}\tau^r_{\mathbf{k}}(E_{\mathbf{k}}-\mu)^s}{1+(\frac{q\tau_{\mathbf{k}}}{m^\star})^2|\mathbf{B}|^2}$$

$$\implies K_{r0} = (\frac{q}{m^\star})^{r-1}\underbrace{\frac{1}{L^d}\sum_{\mathbf{k}}\frac{v^2_{\mathbf{k}}\tau^r_{\mathbf{k}}(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}})}{1+(\frac{q\tau_{\mathbf{k}}}{m^\star})^2|\mathbf{B}|^2}}_{\text{define}=\frac{n_d\langle\tau^r\rangle}{m^\star}} = \frac{q^{r-1}n_d\langle\tau^r\rangle}{(m^\star)^r}. \tag{25.10}$$

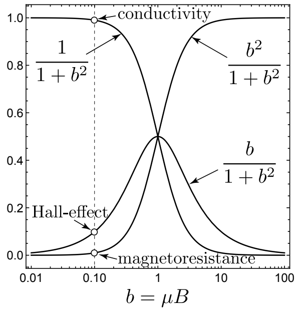


Fig. 25.9 Coefficients of $\mathbf{G}$ vector highlighting the relative importance of the three terms. For low $b^2 = (\mu B)^2 << 1$, the conductivity term is dominant, followed by the Hall effect, and the magnetoresistance term is the smallest. At $b = 1$ they become equal, and for $b^2 >> 1$ their relative importance are switched. This dependence helps make appropriate approximations in analytical calculations of Hall-coefficients and magnetoresistance at low and high magnetic fields.

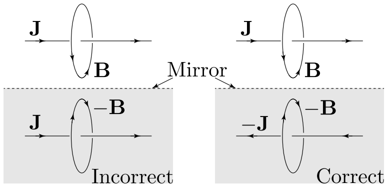


Fig. 25.10 Proof that magnetic field $\mathbf{B}$ is an axial vector (or a pseudovector). A pseudovector has a curious property of mirror reflection that differs strikingly from a polar vector (or real vector) such as an electric field $\mathbf{E}$ or a current $\mathbf{J}$. A polar vector has an ordinary mirror reflection. But an axial vector such as $\mathbf{B}$ must reverse both its direction to $-\mathbf{B}$ *and the direction of its source* to $-\mathbf{J}$ to remain physically correct, as seen in this figure. An axial vector such as the magnetic field (angular momentum is another) is a cross product of two polar vectors. It is therefore constrained; for example, $\nabla\cdot\mathbf{B} = 0$, whereas there is no restriction of this kind for a polar vector.

The definition of the "moments" of the scattering time $\langle \tau^r \rangle$ in the form above may be seen for the simple case of a low magnetic field $B \approx 0$ to be identical to that used for electron transport and mobility in Chapters 21 and 23. In fact, at $\mathbf{B} = 0$ and $\nabla_r T = 0$, Equation 25.7 and 25.10 gives $\mathbf{J} = qK_{10}(q\mathbf{E}) = q^2 K_{10}\mathbf{E} = \sigma\mathbf{E}$ and therefore $\sigma = q^2 K_{10} = q^2 \cdot \frac{n_d \langle \tau \rangle}{m^\star}$, which is the familiar expression for electrical conductivity.

Now consider the current as a function of $\mathbf{B}$ with no concentration gradients, so that $\mathbf{F} = q\mathbf{E}$. From Equation 25.7, we first retain only the normal conductivity and Hall effect portions of the current and leave other components for later discussion. This is certainly justified for $(\mu B)^2 << 1$ as shown in Fig. 25.9. The current density vector $\mathbf{J} = qK_{10}\mathbf{F} + qK_{20}(\mathbf{F} \times \mathbf{B})$ is

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = q^2 K_{10} \begin{bmatrix} E_x \\ E_y \\ 0 \end{bmatrix} + q^2 K_{20} \text{Det} \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ E_x & E_y & 0 \\ 0 & 0 & B_z \end{bmatrix} = q^2 \begin{bmatrix} K_{10}E_x + K_{20}E_y B_z \\ K_{10}E_y - K_{20}E_x B_z \\ 0 \end{bmatrix}, \tag{25.11}$$

on which we can enforce the boundary condition of a finite sample (see Fig. 25.7) at steady state that $J_x = 0$. This immediately yields

$$K_{10}E_x = -K_{20}E_y B_z \implies \frac{E_x}{E_y} = \tan\theta_H = -\frac{K_{20}}{K_{10}} B_z = -\frac{q}{m^\star}\frac{\langle \tau^2 \rangle}{\langle \tau \rangle} B_z$$
$$\implies \tan\theta_H = -\underbrace{\frac{q\langle \tau \rangle}{m^\star}}_{\mu_d} \underbrace{\frac{\langle \tau^2 \rangle}{\langle \tau \rangle^2}}_{r_H} B_z = -\underbrace{r_H \mu_d}_{\mu_H} B_z = -\mu_H B_z. \tag{25.12}$$

Thus from the BTE, we realize that the Hall effect measurement actually yields a Hall angle θ_H that depends via $\tan\theta_H = \mu_H B$ on the **Hall effect mobility** μ_H, and *not* the **drift mobility** μ_d, which is the mobility in the *absence* of a magnetic field[11]. They are related via $\mu_H = r_H \mu_d$, where the **Hall factor** $r_H = \langle \tau^2 \rangle / \langle \tau \rangle^2 = K_{20}/(K_{10}\mu_d)$ is, as shown shortly, of order unity but typically ≥ 1, depending on the energy or momentum dependence of the scattering mechanism via $\tau_{\mathbf{k}}$. For this reason, the Hall effect mobility is equal to, or larger than the drift mobility. Since the drift mobility is what is actually useful in semiconductor devices operating in the absence of an external magnetic field, the difference between μ_H and μ_d should be considered in device design. The calculated mobilities in Chapter 23 were all drift mobilities, and when compared with measured μ_H, should be scaled with the r_H appropriate to the scattering mechanisms.

At small magnetic fields[12] $(\mu B)^2 << 1$, an analytical expression for all dimensions is possible for the Hall factor $r_H = K_{20}/(K_{10}\mu_d)$ if the scattering time can be expressed as $\tau_{\mathbf{k}} = \tau_0 (E_{\mathbf{k}}/k_b T)^p$, where $E_{\mathbf{k}} = \frac{d}{2} m^\star v_k^2$ is the kinetic energy in d-dimensions. The total energy of electrons in the conduction band is $E = E_c + E_k$. Writing $\eta_F = (E_F - E_c)/(k_b T)$, we denote the Fermi–Dirac integral of order j as $F_j(\eta_F)$. Using Equation 21.63 as the generalized expression for drift

[11] The negative sign in Equation 25.12 merely reflects the fact that compared to the infinite sample case, the rotation of the electric field with respect to the current density in a finite sample is in the opposite direction.

[12] An important experimental point here is that the conditions $(\mu B)^2 << 1$ or $(\mu B)^2 >> 1$ are easier to meet than $\mu B << 1$ or $\mu B >> 1$. Several simplifications in semi-classical magnetotransport depend on the criterion $(\mu B)^2 << 1$. For quantum magnetotransport oscillations $\mu B >> 1$ is needed, which requires high magnetic fields and high mobilities.

mobility for parabolic bandstructure, and using integration by parts for the kinetic coefficients (see Exercise 25.2), we get for transport in d-dimensions:

$$K_{r0} \approx K_{r0}(B \to 0) = \left(\frac{q}{m^\star}\right)^{r-1} \frac{1}{L^d} \sum_{\mathbf{k}} \left(-\frac{\partial f^0_{\mathbf{k}}}{\partial E_{\mathbf{k}}}\right) v^2_{\mathbf{k}} \tau^r_{\mathbf{k}}$$

$$= \left(\frac{q}{m^\star}\right)^{r-1} \frac{2\tau_0^r}{dm^\star} \frac{g_s g_v}{(2\pi)^d \Gamma(\frac{d}{2})} \left(\frac{2m^\star k_b T}{\hbar^2}\right)^{\frac{d}{2}} \Gamma(p \cdot r + \frac{d}{2} + 1) F_{p \cdot r + \frac{d}{2} - 1}(\eta_F)$$

$$\implies r_H = \frac{K_{20}}{K_{10}\mu_d} = \frac{\Gamma(2p + \frac{d}{2} + 1)\Gamma(\frac{d}{2} + 1)}{\Gamma^2(p + \frac{d}{2} + 1)} \frac{F_{2p + \frac{d}{2} - 1}(\eta_F) F_{\frac{d}{2} - 1}(\eta_F)}{F^2_{p + \frac{d}{2} - 1}(\eta_F)}$$

$$\text{for } \eta_F << -1, F_j(\eta_F) \approx e^{\eta_F} \text{ and } \boxed{r_H = \frac{\Gamma(2p + \frac{d}{2} + 1)\Gamma(\frac{d}{2} + 1)}{\Gamma^2(p + \frac{d}{2} + 1)}},$$

$$\text{and for } \eta_F >> +1, F_j(\eta_F) \approx \frac{\eta_F^{j+1}}{\Gamma(j+2)} \implies \text{for all } d, p: \boxed{r_H = 1.} \tag{25.13}$$

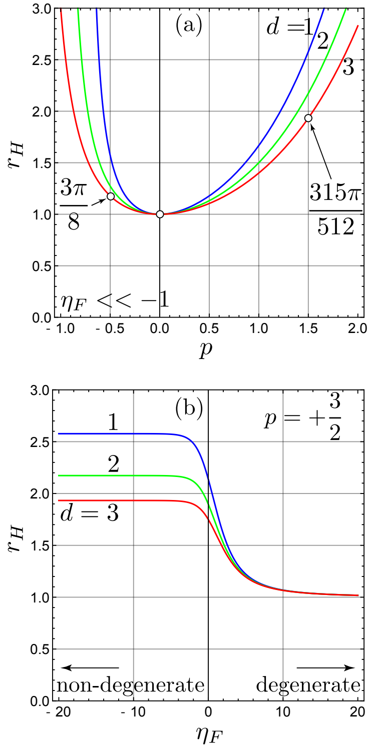


Fig. 25.11 Hall factors $r_H = \mu_H/\mu_d$ for transport in $d = 1, 2, 3$ dimensions in parabolic bands if the scattering time can be expressed as $\tau = \tau_0 (E/k_b T)^p$ where E is the kinetic energy of the carrier. Note the values are typically larger than unity for non-degenerate carriers, the exact value depending on the dimension d and the nature of scattering embodied by p. For degenerate carriers for whom $\eta_F >> +1$, the Hall factor reduces to unity for all cases.

Fig. 25.11 shows the Hall factors for several cases of interest. In Chapter 21 we saw for 3D non-degenerate carriers ($d = 3, \eta_F << -1$), the scattering time by acoustic phonons has $p = -1/2$, which yields $r_H = 3\pi/8 \approx 1.18$. For ionized impurity scattering under same conditions, $p = +3/2$, and $r_H = 315\pi/512 \approx 1.93$. Thus, for non-degenerate ionized impurity scattering, the drift mobility is about half of the Hall effect mobility. Note that for degenerate carrier distributions when $\eta_F >> +1$, the Hall-factor tends to unity for all dimensions for all scattering events. Therefore for degenerate carrier concentrations, the Hall effect mobility is the same as the field-effect mobility.

Ambipolar and multi-carrier transport: Consider the case when there are multiple types of carriers that are responsible for current flow. These could be, for example, carriers in the light-hole and heavy-hole bands in a p-type semiconductor, or electrons in the conduction band and holes in the valence bands. If we assume the carriers of different types do not affect each other, the Hall effect measurement yields a carrier density and mobility that is a mixture of the two types. Consider the situation when n electrons of mobility μ_n and p holes of mobility μ_p are simultaneously involved in transport. In that case, the net charge current is given by (retaining only the conductivity and Hall terms and neglecting the others):

$$\mathbf{J}_n = \sigma_n \mathbf{E} - \mu_n (\mathbf{J}_n \times \mathbf{B}) \implies \mathbf{J}_n = \frac{nq\mu_n \mathbf{E} - nq\mu_n^2 (\mathbf{E} \times \mathbf{B}) + nq\mu_n^3 (\mathbf{E} \cdot \mathbf{B})\mathbf{B}}{1 + (\mu_n B)^2},$$

$$\mathbf{J}_p = \sigma_p \mathbf{E} + \mu_p (\mathbf{J}_p \times \mathbf{B}) \implies \mathbf{J}_p = \frac{pq\mu_p \mathbf{E} + pq\mu_p^2 (\mathbf{E} \times \mathbf{B}) + pq\mu_p^3 (\mathbf{E} \cdot \mathbf{B})\mathbf{B}}{1 + (\mu_p B)^2}. \tag{25.14}$$

Now using $\mathbf{J} = \mathbf{J}_n + \mathbf{J}_p$ and working through the same procedure as Equations 25.11 and 25.12, enforcing $J_x = 0$ and writing $b = \mu_n/\mu_p$, the mixed magnetic field-dependent Hall-conductivity and Hall-coefficient are obtained as

$$\sigma_{xx} = \frac{J_x}{E_x} = \frac{q[(n\mu_n + p\mu_p)^2 + (n-p)^2\mu_n^2\mu_p^2 B^2]}{(n\mu_n + p\mu_p) + (n\mu_p + p\mu_n)\mu_n\mu_p B^2}$$

$$R_H = \frac{1}{\sigma_{xy}B} = \frac{E_y}{BJ_x} = \frac{(p - nb^2) + (p-n)b^2\mu_p^2 B^2}{q[(p+nb)^2 + (p-n)^2 b^2 \mu_p^2 B^2]}. \tag{25.15}$$

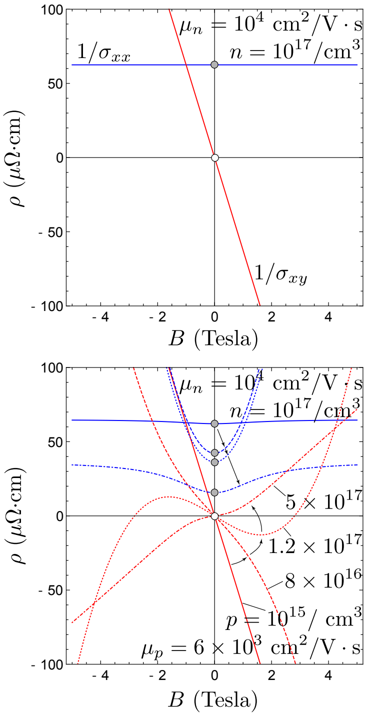


Fig. 25.12 Ambipolar Hall effect. (a) Hall effect resistivities for single n-type carriers. Note the zero crossings, and the negative slope of $1/\sigma_{xy} = R_H B$. (b) How the Hall effect resistivities change as the hole density increases from $p = 10^{15}/\text{cm}^3$ which is far lower than n-type carriers, to $p = 5 \times 10^{17}$ $/\text{cm}^3$ which is larger. $1/\sigma_{xy} = R_H B$ changes from negative to positive slope, and $1/\sigma_{xx}$ shows dependence on B, which is weak for $n >> p$ or $p >> n$, and is the strongest when $n \approx p$. The end-of chapter references discuss multi-carrier transport analysis further.

Fig. 25.12 (a) shows $\frac{1}{\sigma_{xx}}$ and $\frac{1}{\sigma_{xy}} = R_H B$ as a function of the magnetic field B for $n = 10^{17}/\text{cm}^3$ and $\mu_n = 10^4$ cm^2/V·s and $p = 0$, i.e., for transport by carriers of one type. The negative slope of the Hall resistance R_H vs. B indicates that the mobile carriers are negatively charged. $\frac{1}{\sigma_{xx}}$ does not vary appreciably with B for this case, indicating no magnetoresistance. But we shall have more to discuss about magnetoresistance in the next section. Fig. 25.12 (b) shows how the dependence of $\frac{1}{\sigma_{xx}}$ and $R_H = \frac{1}{B\sigma_{xy}}$ evolves as the carriers involved in transport become increasingly ambipolar, from predominantly n-type, to mixed n- and p-type, to predominantly p-type. The Hall coefficient R_H dependence on B switches from a negative to positive slope for the extreme cases, going through zero-crossings in the ambipolar case. At the same time, ambipolar, or multi-carrier transport results in the appearance of the dependence of $\frac{1}{\sigma_{xx}}$ on B. Large dependence of magnetoresistance effects in σ_{xx} in several cases signal the **presence** of carriers of different signs, or of the same sign but different effective masses and mobilities. These features are turned around and used to identify multiple carrier species and of various signs from the experiment. The next section discusses intrinsic magnetoresistance.

25.3 Magnetoresistance

When $\mathbf{B} \perp \mathbf{E}$ for an infinite sample as shown in Figures 25.7 (a, c), since the magnetic field curves the path of electrons away from the y-axis between every successive scattering event, the current in the y-direction is reduced compared to the case when there is no magnetic field. Thus, one should expect a positive magnetoresistance. The current in the direction of the electric field is $J_p \cos\theta_H$. A crude estimate may be obtained for the case when $\mu B << 1$ in which case $\tan\theta_H = \mu B << 1$. Then, $\tan\theta_H \approx \theta_H = \mu B$, and $\cos\theta_H \approx 1 - \theta_H^2/2 = 1 - (\mu B)^2/2$. Thus, the magnetoresistance is $\frac{\Delta R}{R} \approx \frac{J_p(B=0)}{J_p(B) - J_p(B=0)} \approx 1 + \frac{(\mu B)^2}{2}$, and the magnetoresistance is roughly a factor $(\mu B)^2/2$ of the resistance in the absence of the magnetic field.

The magnetoresistance of a finite semiconductor sample also increases roughly as $(\mu B)^2$, up to an additional factor of order unity. This is typically small, except for high mobility semiconductors such as InSb. It is essential to now include the third term proportional to

$\mathbf{B}(\mathbf{F}\cdot\mathbf{B})$ in Equation 25.7 and also retain the magnetic field in the denominator of the kinetic coefficients of Equation 25.10. We exclude the terms due to temperature gradients for the next section, and solve for the magnetoresistance in a finite sample again, using the same procedure as the Hall effect, but for a general **E** and **B** that could point in any direction.

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = q^2 \times$$

$$\left(K_{10} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} + K_{20}\mathrm{Det} \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ E_x & E_y & E_z \\ B_x & B_y & B_z \end{bmatrix} + K_{30} \begin{bmatrix} (B_xE_x + B_yE_y + B_zE_z)B_x \\ (B_xE_x + B_yE_y + B_zE_z)B_y \\ (B_xE_x + B_yE_y + B_zE_z)B_z \end{bmatrix} \right)$$

$$= q^2 \begin{bmatrix} K_{10}E_x + K_{20}(E_yB_z - E_zB_y) + K_{30}(B_xE_x + B_yE_y + B_zE_z)B_x \\ K_{10}E_y + K_{20}(E_zB_x - E_xB_z) + K_{30}(B_xE_x + B_yE_y + B_zE_z)B_y \\ K_{10}E_z + K_{20}(E_xB_y - E_yB_x) + K_{30}(B_xE_x + B_yE_y + B_zE_z)B_z \end{bmatrix}$$

$$= q^2 \begin{bmatrix} K_{10} + K_{30}B_x^2 & K_{20}B_z + K_{30}B_xB_y & -K_{20}B_y + K_{30}B_xB_z \\ -K_{20}B_z + K_{30}B_xB_y & K_{10} + K_{30}B_y^2 & K_{20}B_x + K_{30}B_yB_z \\ K_{20}B_y + K_{30}B_xB_z & -K_{20}B_x + K_{30}B_yB_z & K_{10} + K_{30}B_z^2 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}$$

$$\text{For } B_x = B_y = 0, \quad \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = q^2 \begin{bmatrix} K_{10} & K_{20}B_z & 0 \\ -K_{20}B_z & K_{10} & 0 \\ 0 & 0 & K_{10} + K_{30}B_z^2 \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}$$

$$\implies \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = \frac{1}{q^2} \begin{bmatrix} \frac{K_{10}}{K_{10}^2 + K_{20}^2B_z^2} & -\frac{K_{20}B_z}{K_{10}^2 + K_{20}^2B_z^2} & 0 \\ \frac{K_{20}B_z}{K_{10}^2 + K_{20}^2B_z^2} & \frac{K_{10}}{K_{10}^2 + K_{20}^2B_z^2} & 0 \\ 0 & 0 & \frac{1}{K_{10} + K_{30}B_z^2} \end{bmatrix} \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix}. \tag{25.16}$$

From Equation 25.12, since $(K_{20}/K_{10})^2 = \mu_H^2$, the kinetic coefficients K_{r0} depend on the magnetic field via the $1 + (\omega_c\tau_{\mathbf{k}})^2$ terms in the denominators. For small magnetic fields $\frac{1}{1+(\omega_c\tau_{\mathbf{k}})^2} \approx 1 - \omega_c^2\tau_k^2$. Denoting $\langle ... \rangle$ as averaging at $B = 0$ and retaining terms up to order B_z^2, we get[13]

$$K_{r0} = \left(\frac{q}{m^\star}\right)^{r-1} \frac{1}{L^d} \sum_{\mathbf{k}} \frac{v_{\mathbf{k}}^2 \tau_{\mathbf{k}}^r \left(-\frac{\partial f_{\mathbf{k}}^0}{\partial E_{\mathbf{k}}}\right)}{1 + \left(\frac{q\tau_{\mathbf{k}}}{m^\star}\right)^2 |\mathbf{B}|^2} \approx \frac{q^{r-1} n_d}{(m^\star)^r} \times \left(\langle \tau^r \rangle - \omega_c^2 \langle \tau^{r+2} \rangle \right),$$

$$\implies \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} =$$

$$\frac{m^\star}{n_d q^2 \langle\tau\rangle} \begin{bmatrix} 1 + \frac{\langle\tau^3\rangle\langle\tau\rangle - \langle\tau^2\rangle^2}{\langle\tau\rangle^4} \left(\frac{q\langle\tau\rangle}{m^\star} B_z\right)^2 & -\frac{q\langle\tau^2\rangle}{m^\star\langle\tau\rangle} B_z & 0 \\ \frac{q\langle\tau^2\rangle}{m^\star\langle\tau\rangle} B_z & 1 + \frac{\langle\tau^3\rangle\langle\tau\rangle - \langle\tau^2\rangle^2}{\langle\tau\rangle^4} \left(\frac{q\langle\tau\rangle}{m^\star} B_z\right)^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix}. \tag{25.17}$$

Equation 25.17 is in the form $\mathbf{E} = \rho\mathbf{J}$ where ρ, the resistivity tensor is a product of $\rho_0 = \frac{m^\star}{n_d q^2 \langle\tau\rangle}$, the resistivity at zero magnetic field and a matrix whose deviation from the unit matrix accounts for the effects of the magnetic field B_z. We again recognize the transverse resistivity for $J_x = 0$ to be $\rho_{xy} = E_x/J_y = -\frac{q\langle\tau^2\rangle}{m^\star\langle\tau\rangle}B_z = \mu_H B_z$ as the Hall resistance. The longitudinal magnetoresistance is therefore seen to be

---

[13] Written out explicitly for $r = 1, 2,$ and 3, $K_{10} = \frac{n_d(\langle\tau\rangle - \omega_c^2\langle\tau^3\rangle)}{m^\star}$, $K_{20} = \frac{q n_d(\langle\tau^2\rangle - \omega_c^2\langle\tau^4\rangle)}{(m^\star)^2}$, and $K_{30} = \frac{q^2 n_d(\langle\tau^3\rangle - \omega_c^2\langle\tau^5\rangle)}{(m^\star)^3}$.

$$\frac{\Delta\rho_{yy}}{\rho_{yy}} = \left(\frac{\langle\tau^3\rangle}{\langle\tau\rangle^3} - \frac{\langle\tau^2\rangle^2}{\langle\tau\rangle^4}\right)(\mu_d B)^2 = M_R(\mu_d B)^2. \tag{25.18}$$

As anticipated from simpler arguments, the magnetoresistance is therefore proportional to $(\mu_d B_z)^2$ where μ_d is the drift mobility: a higher mobility leads to a higher magnetoresistance. For the circumstances discussed here, the dimensionless term M_R in brackets due to the scattering times is positive (try proving it!). If the scattering time can be expressed as $\tau_{\mathbf{k}} = \tau_0(E_{\mathbf{k}}/k_bT)^p$, as before, we get

$$M_R = \frac{\langle\tau^3\rangle}{\langle\tau\rangle^3} - \frac{\langle\tau^2\rangle^2}{\langle\tau\rangle^4} = \frac{\Gamma(\frac{d}{2}+1)^2 F^2_{\frac{d}{2}-1}(\eta_F)}{\Gamma^4(p+\frac{d}{2}+1)F^4_{p+\frac{d}{2}-1}(\eta_F)} \times$$
$$\left(\Gamma(3p+\frac{d}{2}+1)\Gamma(p+\frac{d}{2}+1)F_{3p+\frac{d}{2}-1}(\eta_F)F_{p+\frac{d}{2}-1}(\eta_F) - \Gamma^2(2p+\frac{d}{2}-1)F^2_{2p+\frac{d}{2}-1}(\eta_F)\right).$$
$$\text{For } \eta_F << -1, d = 3, M_R = \frac{9\pi}{16}\cdot\frac{\Gamma(3r+\frac{5}{2})\Gamma(r+\frac{5}{2}) - \Gamma^2(2r+\frac{5}{2})}{\Gamma^4(3r+\frac{5}{2})},$$
$$\text{and for } \eta_F >> +1, M_R = 0 \text{ for all } d. \tag{25.19}$$

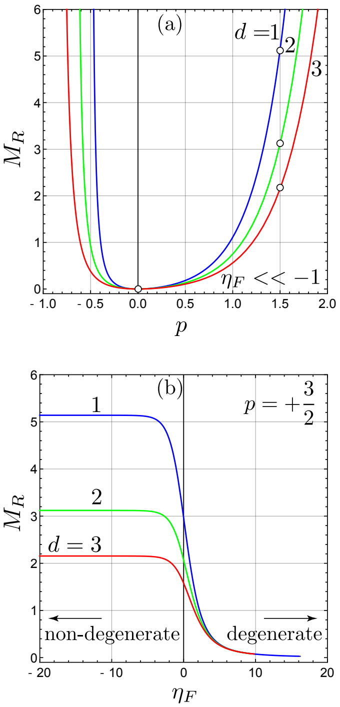


Fig. 25.13 Magnetoresistance factors.

Fig. 25.13 (a) shows the dependence of the magnetoresistance factor M_R for non-degenerate carrier transport in various dimensions, indicating a larger variation than the Hall factor. Note that if the scattering time does not depend on the carrier energy ($p = 0$), the magnetoresistance goes to zero. Fig. 25.13 (b) shows the dependence of M_R on the level of degeneracy, where it is clearly seen that magnetoresistance vanishes for degenerate carrier distributions. Thus, if the momentum scattering time $\tau_{\mathbf{k}}$ is independent of the energy via the bandstructure $E_{\mathbf{k}}$, or if all carriers responsible for transport have the same energy (as is the case for metals or degenerate semiconductors), the magnetoresistance is zero.

It must be kept in mind that we have assumed parabolic bandstructure and low magnetic fields to arrive at these results. Exercise 25.3 discusses cases when these assumptions are removed. Of historical importance is the fact that angle-dependent magnetoresistance measurements provided the first experimental evidence of the anisotropic conduction bandstructures and multi-valley minima of silicon and germanium. The advent of cyclotron resonance provided more accurate transport methods to probe bandstructure, as discussed later in this chapter.

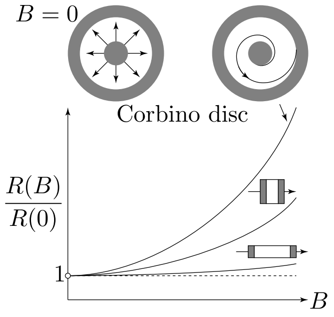


Fig. 25.14 Increase in magnetoresistance due to shape of semiconductor. The magnetic field points into the plane of the paper. The magnetoresistance of a long and thin semiconductor is low: a wide and short semiconductor has a higher magnetoresistance. Because of the out-of-plane magnetic field, the current trajectories spiral from the inner to outer contacts of the Corbino disc, causing a large increase in the magnetoresistance.

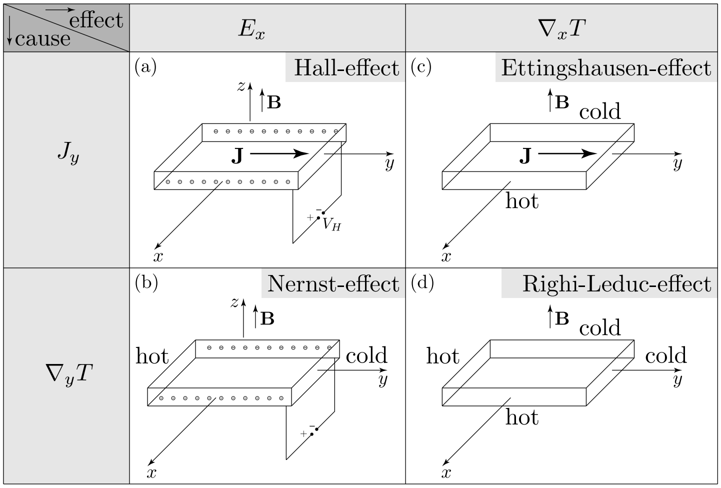


Fig. 25.15 Four magnetotransport effects resulting from currents and temperature gradients for a magnetic field along z-direction. (a) Hall effect, (b) Nernst effect, (c) Ettingshausen effect, and (d) Righi–Leduc effect.

A large magnetoresistance is attractive as a sensor of magnetic fields. Magnetoresistance effects as well as Hall effect are widely used for the fabrication of sensors. The **geometry** of the semiconductors have a strong effect on the magnetoresistance, as may be expected. From an intuitive picture, instead of a linear Hall effect bar, a circular ring of contacts called **Corbino discs** are used for enhancing magnetoresistance[14], as shown in Fig. 25.14. The current that flowed radially outward in the absence of a magnetic field now follows "spiral" trajectories as shown. This increases the path length significantly, and therefore causes a large increase in the magnetoresistance.

25.4 Nernst and Ettingshausen effects

Fig. 25.15 shows additional magnetotransport effects that result from **temperature gradients** due to a magnetic field in the z-direction. Apart from the Hall effect, the strongest is the **Nernst effect** shown in Fig. 25.15 (b), in which a temperature gradient $\nabla_y T$ along the y-axis causes an electric field E_x, and hence a voltage to develop in the x-axis. It is the thermal analogue of the Hall effect. In the **Ettingshausen effect** of Fig. 25.15 (c), a current J_y in the y-direction causes a temperature gradient $\nabla_x T$ to develop in the x-direction[15]. In the **Righi–Leduc** effect of Fig. 25.15 (d), a temperature gradient along the y-axis causes a temperature gradient along the x-axis. Compared to the Nernst effect in which a voltage difference is measured, the Ettingshausen and Righi–Leduc effects are rather difficult to measure because $\Delta_x T$ in semiconductors is only weakly determined by electrons and holes but significantly by *phonon* conduction.

We now discuss the calculation of the Nernst coefficient as a template for the calculation of all other thermo-magnetic coefficients. The starting point again is the BTE solution in Equation 25.7. We neglect

[14] Corbino-disc geometry in the MOSFET structure was the first used to observe Shubnikov de Haas oscillations in 2DEGs as described in Physical Review Letters, vol. 16. page 901 (1966). In fact the precursor to the quantum Hall effect, indicated by a vanishing R_{xx} was also observed in the Corbino disk, as described in the Journal of the Physical Society of Japan, vol. 44, page 1839 (1978).

[15] The Ettingshausen effect typically occurs in parallel with the Hall effect, and therefore accounts for "errors" in the Hall effect measurement, albeit of a small magnitude. Since it depends upon the sign of the magnetic field, the error is eliminated if the Hall effect measurement is performed twice with the magnetic field pointing along the $+z$ and $-z$ directions, and the average Hall voltage is taken.

terms in B^2 for $\mathbf{B} = B_z\hat{z}$ to write the charge current density as

$$\mathbf{J} = Q^2 K_{10}\mathbf{E} + Q^2 K_{20}(\mathbf{E}\times\mathbf{B}) - \frac{Q}{T}K_{11}\nabla_{\mathbf{r}}T - \frac{Q}{T}K_{21}(\nabla_{\mathbf{r}}T\times\mathbf{B})$$

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} =$$

$$Q^2\begin{bmatrix} K_{10} & K_{20}B_z & 0 \\ -K_{20}B_z & K_{10} & 0 \\ 0 & 0 & K_{10} \end{bmatrix}\begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} - \frac{Q}{T}\begin{bmatrix} K_{11} & K_{21}B_z & 0 \\ -K_{21}B_z & K_{11} & 0 \\ 0 & 0 & K_{11} \end{bmatrix}\begin{bmatrix} \nabla_x T \\ \nabla_y T \\ \nabla_z T \end{bmatrix}$$

$$\text{for } \mathbf{J} = 0 \implies$$

$$\begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = \frac{1}{QT}\begin{bmatrix} -\frac{K_{10}K_{11}+K_{20}K_{21}B_z^2}{K_{10}^2+K_{20}^2B_z^2} & \frac{K_{11}K_{20}-K_{10}K_{21}}{K_{10}^2+K_{20}^2B_z^2}B_z & 0 \\ -\frac{K_{11}K_{20}-K_{10}K_{21}}{K_{10}^2+K_{20}^2B_z^2}B_z & -\frac{K_{10}K_{11}+K_{20}K_{21}B_z^2}{K_{10}^2+K_{20}^2B_z^2} & 0 \\ 0 & 0 & -\frac{K_{11}}{K_{10}} \end{bmatrix}\begin{bmatrix} \nabla_x T \\ \nabla_y T \\ \nabla_z T \end{bmatrix}$$

$$M_N = \frac{E_x}{(-\nabla_y T)B_z} = -\frac{1}{QT}\frac{K_{11}K_{20} - K_{10}K_{21}}{K_{10}^2 + K_{20}^2 B_z^2} \approx \frac{1}{QT}\left(\frac{K_{10}K_{21} - K_{11}K_{20}}{K_{10}^2}\right). \tag{25.20}$$

The Nernst coefficient M_N can be expressed in terms of the carrier mobilities and the ratios of the scattering times by substitution of the kinetic coefficients K_{rs} as before. The Nernst coefficient is of the order $M_N \approx \frac{k_b}{q}\cdot\mu_H p$, where μ_H is the Hall effect mobility, and p is the power dependence of the scattering time $\tau_{\mathbf{k}}$. A typical voltage difference for $\nabla_y T = 10$ K/cm and $l_x = 1$ mm is $E_x l_x \approx 1\ \mu$V.

For the calculation of the Ettingshausen and the Righi–Leduc coefficients, a similar procedure is followed, but starting from the energy transport Equation 25.9. Onsager relations connect the various coefficients. Experimental measurements of the Hall, Nernst, Ettingshausen, and Righi–Leduc effects were historically influential in the formulation of the fundamental laws of thermodynamics, entropy, and symmetry of physical properties embodied in the Onsager relations discussed in Section 21.10. Adiabatic and isothermal aspects of these effects therefore should be treated separately for accurate comparison with experiments.

Thermo-magnetic effects play important roles in spintronic devices such as spin-transfer torque (STT) or spin-orbit torque (SOT) magnetic random-access memories (MRAMs) made of magnetic materials and topological insulators. In the rest of the chapter, we neglect temperature gradients, and focus primarily on the effect of magnetic and electric fields in semiconductors.

25.5 Cyclotron resonance

In the past sections we have discussed steady-state, or time-independent magnetotransport phenomena. The influence of time-varying fields and currents in the presence of a DC magnetic field holds special historical significance in the uncovering of the details of semiconductor bandstructure and measurements of effective masses of electrons and holes. The prime example of this effect is **cyclotron resonance**, which occurs when an oscillating electric field interacts with mobile carriers in the presence of a DC magnetic field. Cyclotron resonance was among the first experimental probes of semiconductor bandstructures and their anisotropies.

The physical mechanism is the following: in a magnetic field, an electron undergoes cyclotron orbits of frequency $\omega_c = qB/m^\star$. If an electromagnetic field of circular frequency ω is incident on the semiconductor, there is a strong absorption when $\omega = \omega_c$. The absorption peak (see Fig. 25.16) is thus a direct measure of the effective mass. By changing the frequency and the direction of the magnetic field, details of the bandstructure of the occupied states are obtained.

For time-varying fields, in Chapter 21, Equation 21.83, by using the moments of the BTE, we obtained the following relation for the current density $\mathbf{J}$ in response to electric and magnetic fields and concentration gradients:

$$\boxed{\frac{\partial \mathbf{J}_e}{\partial t} + \frac{\mathbf{J}_e}{\tau_e} = \frac{nq^2}{m_c^\star}\mathbf{E} - \frac{q}{m_c^\star}(\mathbf{J}_e \times \mathbf{B}) + \frac{qD_e}{\tau_e}\nabla_{\mathbf{r}} n,} \tag{25.21}$$

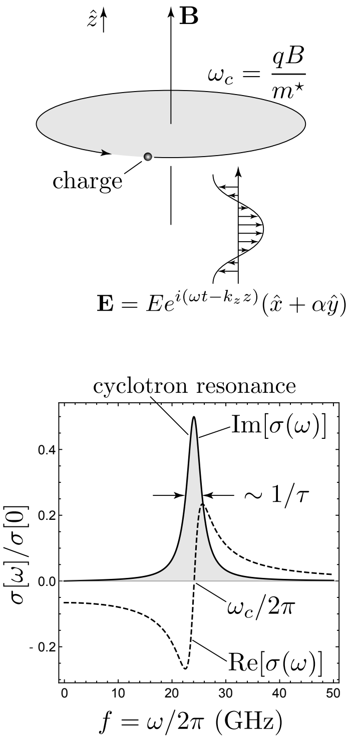


Fig. 25.16 Cyclotron resonance and its application in measurement of semiconductor bandstructure parameters. The typical cyclotron resonance frequencies are in the 10s of GHz, and the width of the resonance peak is inversely proportional to the scattering time and mobility of carriers.

where the subscript e is added for electrons. For holes, the signs of the terms linear in q flip sign:

$$\boxed{\frac{\partial \mathbf{J}_h}{\partial t} + \frac{\mathbf{J}_h}{\tau_h} = \frac{pq^2}{m_v^\star}\mathbf{E} + \frac{q}{m_v^\star}(\mathbf{J}_h \times \mathbf{B}) - \frac{qD_h}{\tau_h}\nabla_{\mathbf{r}} p.} \tag{25.22}$$

In the above relations, τ_e and τ_h are the momentum scattering rates of electrons and holes respectively of effective masses $m_c^\star$ and $m_v^\star$, and diffusion constants D_e and D_v. These two relations are central to understanding all time-dependent magnetotransport phenomena such as cyclotron resonance (this section), and Faraday rotation (next section).

A homogeneously doped n-type semiconductor ($\nabla_r n = 0$) is subjected to a DC magnetic field $\mathbf{B} = B\hat{z}$, and an electromagnetic wave of frequency ω propagating along the z-direction (Fig. 25.16) with the electric field component $\mathbf{E} = Ee^{i\omega t}(\hat{x} + \alpha\hat{y})$. The electromagnetic wave is linearly polarized if $\alpha = 0$, and circularly polarized for $\alpha = \pm i$. The magnetic component of the electromagnetic field is much smaller than the DC magnetic field $\mathbf{B} = B\hat{z}$ and is therefore neglected in the following analysis. Assuming mobile electrons respond to the wave with a current $\mathbf{J} = e^{i\omega t}(J_x\hat{x} + J_y\hat{y} + J_z\hat{z})$, from Equation 25.21 for isotropic bandstructure with $\omega_c = qB/m^\star$ and DC conductivity $\sigma_0 = nq^2\tau/m^\star$,

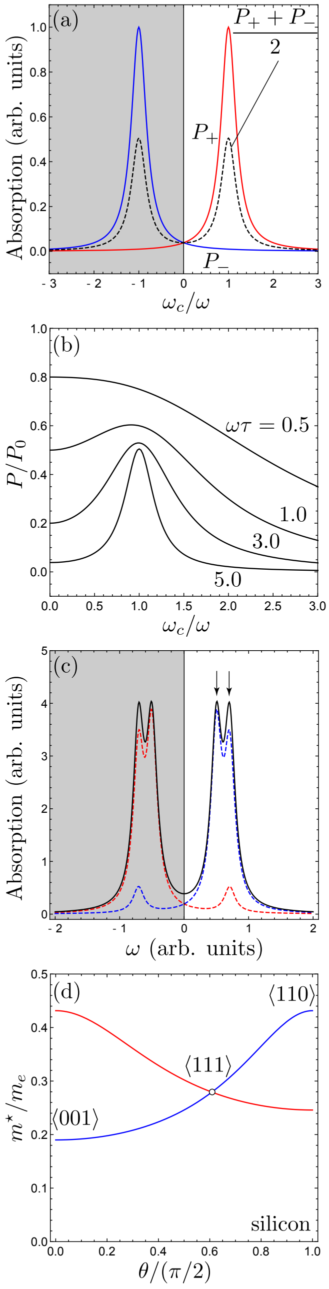


Fig. 25.17 (a) Cyclotron resonance absorption for right, left, and plane polarized microwave radiation. (b) Absorption for various $\omega\tau$. (c) Multiple peaks for anisotropic effective masses, and (d) Effective masses of silicon vs. angle of crystal with **B** from cyclotron resonance.

the conductivity tensor is obtained as a function of ω:

$$i\omega\tau\begin{bmatrix}J_x\\J_y\\J_z\end{bmatrix}e^{i\omega t}+\begin{bmatrix}J_x\\J_y\\J_z\end{bmatrix}e^{i\omega t}=nq\mu E\begin{bmatrix}1\\\alpha\\0\end{bmatrix}e^{i\omega t}-\mu\text{Det}\begin{bmatrix}\hat{x}&\hat{y}&\hat{z}\\J_x&J_y&J_z\\0&0&B\end{bmatrix}e^{i\omega t}$$

$$\begin{bmatrix}1+i\omega\tau&\mu B&0\\-\mu B&1+i\omega\tau&0\\0&0&1+i\omega\tau\end{bmatrix}\begin{bmatrix}J_x\\J_y\\J_z\end{bmatrix}=qn\mu\begin{bmatrix}E_x\\E_y\\E_z\end{bmatrix}$$

$$\implies\begin{bmatrix}J_x\\J_y\\J_z\end{bmatrix}=\sigma_0\underbrace{\begin{bmatrix}\frac{1+i\omega\tau}{(\omega_c\tau)^2+(1+i\omega\tau)^2}&-\frac{\omega_c\tau}{(\omega_c\tau)^2+(1+i\omega\tau)^2}&0\\\frac{\omega_c\tau}{(\omega_c\tau)^2+(1+i\omega\tau)^2}&\frac{1+i\omega\tau}{(\omega_c\tau)^2+(1+i\omega\tau)^2}&0\\0&0&\frac{1}{1+i\omega\tau}\end{bmatrix}}_{\sigma}\begin{bmatrix}E_x\\E_y\\E_z\end{bmatrix}$$

$$\text{for right-circularly polarized } E_y=-iE_x\implies\sigma_+=\frac{J_x}{E_x}=\sigma_{xx}-i\sigma_{yy}$$

$$\text{and absorbed power is } P_+=\frac{1}{2}\text{Re}[\mathbf{J}^T\mathbf{E}^\star]=\frac{1}{2}|\mathbf{E}|^2\text{Re}[\sigma_+]$$

$$\text{for left-circularly polarized } E_y=+iE_x\implies\sigma_-=\frac{J_x}{E_x}=\sigma_{xx}+i\sigma_{yy}$$

$$\text{and absorbed power is } P_-=\frac{1}{2}|\mathbf{E}|^2\text{Re}[\sigma_-]$$

$$\text{and for linearly polarized the absorbed power is } P=\frac{P_++P_-}{2}. \tag{25.23}$$

For isotropic bandstructure we get $\sigma_+/\sigma_0 = 1/(1+i(\omega+\omega_c)\tau)$ and $\sigma_-/\sigma_0 = 1/(1+i(\omega-\omega_c)\tau)$, from where P_+, P_-, P are obtained and plotted in Fig. 25.17 (a). The difference between right and left-circularly polarized light is important: a peak of absorption is observed for the circular polarization which is in the same direction as the cyclotron motion of the charge due to the magnetic field. Therefore **the sign of the charge** is concluded from the dependence of the absorption spectrum on the polarization, and the peak of absorption is a direct measure of the **effective mass** $m^\star$. Fig. 25.17 (b) shows that cyclotron resonance places a stringent requirement of $\omega\tau \geq 1$ for a peak to be clearly resolved. Since the microwave frequencies $f = \omega/2\pi$ used are in the 10s of GHz, the scattering time must be several ps. This is possible at low temperatures in very high purity semiconductors of high free carrier mobility.

Cyclotron resonance can actually do more: it can give the anisotropic effective masses and also information about the number of valleys. Consider the magnetic field to be oriented along an arbitrary direction $\mathbf{B} = B(\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta)$. For ellipsoidal bandstructure

$$E(\mathbf{k})=\frac{\hbar^2}{2}\left(\frac{k_x^2}{m_x}+\frac{k_y^2}{m_y}+\frac{k_z^2}{m_z}\right), \tag{25.24}$$

we rewrite the effective mass $m^\star \to [M]$ as a tensor with ellipsoidal bandstructure to obtain the conductivity tensor:

$$[M]^{-1} = \begin{bmatrix} \frac{1}{m_x} & 0 & 0 \\ 0 & \frac{1}{m_y} & 0 \\ 0 & 0 & \frac{1}{m_z} \end{bmatrix} \implies \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} =$$

$$\underbrace{nq^2\tau \begin{bmatrix} 1+i\omega\tau & \frac{q\tau}{m_x}B\cos\theta & -\frac{q\tau}{m_x}B\sin\phi\sin\theta \\ -\frac{q\tau}{m_y}B\cos\theta & 1+i\omega\tau & \frac{q\tau}{m_y}B\cos\phi\sin\theta \\ \frac{q\tau}{m_x}B\sin\phi\sin\theta & -\frac{q\tau}{m_z}B\cos\phi\sin\theta & 1+i\omega\tau \end{bmatrix}^{-1} [M]^{-1}}_{\sigma} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix}. \tag{25.25}$$

The expression for conductivity obtained after matrix inversion is long but analytical, hence is not written out explicitly. The components of the resulting conductivity matrix are used as in Equation 25.23 to obtain the absorption spectra P_+, P_-, P as a function of ω_c/ω. Fig. 25.17 (c) shows that for $m_x = m_y = m_t \neq m_z = m_l$, there are two peaks in the absorption spectra for plane polarized microwave radiation[16]. The split in the peak frequencies depends on the angle that the crystal (and hence the ellipsoidal bandstructure) makes with the magnetic field. It also depends on the occupation of carriers in multiple valleys, which contribute to the absorption independently if intervalley scattering is neglected.

For silicon, if **B** is in the (110) crystal plane, and makes an angle θ with the $[001]$ axis, the angle dependence of the peak cyclotron frequencies occurs at the effective masses

$$m^\star_{c,1} = \sqrt{\frac{m_l m_t^2}{m_t + (m_l - m_t)\cos^2\theta}} \text{ and } m^\star_{c,2} = \sqrt{\frac{2m_l m_t^2}{2m_t + (m_l - m_t)\sin^2\theta}}, \tag{25.26}$$

from where $m_l = 0.98m_e$ and $m_t = 0.19m_e$ are extracted by performing the measurement as a function of θ. Thus, by changing the angle the effective mass vs. angle is experimentally measured, as shown for the case of silicon[17] in Fig. 25.17 (d).

25.6 Faraday rotation

Consideration of both electric and magnetic components of a propagating linearly polarized electromagnetic wave through a semiconductor in the direction of an external DC magnetic field causes a rotation of the plane of polarization. This effect, called **Faraday rotation**[18] occurs because plane polarized light is a combination of right and left circular polarization, which interact differently with the free carriers undergoing cyclotron motion. Because of the change of refractive index due to the free carriers that each polarization interacts with, the two polarizations experience different path lengths. The analysis of the Faraday rotation proceeds along similar lines as cyclotron resonance,

[16] The red and blue dashed lines are the left and right circularly polarized contributions.

[17] The red and blue lines are for the two masses of Equation 25.26.

[18] The discovery of this effect is rather remarkable. Just like one magnet rotates another, Faraday hypothesized that if light has a magnetic field component in it, the direction of its field polarization should be deflected when going through a material under a strong magnetic field. But to prove it, he had to wait for several years to obtain the discarded anchor of a ship of the Royal Navy to make a strong enough electromagnet. Only with the highest magnetic field he could observe the rotation!

but now both the electric and magnetic components of the electromagnetic wave are introduced into Equation 25.21, and the effective refractive index experienced by the light wave is evaluated for left and right circular polarizations. For a magnetic field B and propagation thickness d, at light frequency ω the Faraday rotation angle θ_F is related to the material parameters as

$$\frac{\theta_F}{Bd} = \frac{nq^3}{2n_r \epsilon_0 c (m^\star)^2 (\omega^2 - \omega_c^2)}. \tag{25.27}$$

Thus Faraday rotation for $\omega >> \omega_c$ can be used to measure the effective mass if the carrier density n and the refractive index n_r at the frequency ω are known. This is useful for measuring $m^\star$ of semiconductors in which the stringent conditions needed for cyclotron resonance are not met. The sense of Faraday rotation (clockwise or counterclockwise) depends on the sign of the charge – allowing the distinction of electrons and holes. It also depends on whether the light wave propagates along or against[19] **B**. As a numerical example, for $m^\star = 0.2 m_e$, $d = 1$ mm, $B = 1$ T, $n = 10^{18}/\text{cm}^3$, $n_r = \sqrt{10}$, and $\lambda = 10$ μm, the angle of Faraday rotation in degrees is $\theta_F = 11.9^\circ$.

If the semiconductor itself is magnetic (e.g. GaMnAs and CdMnTe are dilute magnetic semiconductors), additional rotation of the polarization is observed and can be a very strong effect. Faraday rotation is a rare example of a **non-reciprocal effect** in light wave propagation and therefore is used for making **optical isolators** where light propagation in one direction is desired but not the reverse. The quantity $\theta_F/(Bd)$ is also called the **Verdet constant** of a material, and is the basis of choosing materials for optical isolators. The highest Verdet constants are $\sim$ 100 radians/(Tesla.meter) in terbium gallium garnet (TGG) crystals. Semiconductors have lower Verdet constants of $\sim$ 10s of radians/(Tesla.meter), but can reach larger values for light of photon energy close to the bandgap[20].

25.7 Atoms: Bohr magneton and spin

We now make the transition from semi-classical effects of magnetic fields on transport phenomena in semiconductors to the fully quantum ones. The point of distinction between the two is that the semi-classical effects can be explained by assuming that the electronic bandstructure of the semiconductor is unaffected in the presence of a magnetic field. For large magnetic fields, this is no longer the case. The presence of a large magnetic field changes the energy eigenvalues in a manner that the density of states of a semiconductor is modified significantly. The concept of the wavevector-dependent bandstructure $E(\mathbf{k})$ partially loses its validity because the wavevectors (k_x, k_y) are mixed strongly by a magnetic field along the z-direction due to classical circular motion perpendicular to the magnetic field. In large B fields the energy eigenvalues and eigenfunctions of electrons in a solid change to

---

[19] Faraday rotation is thus seen to be the optical analogue of the Hall effect. It maybe referred to loosely as the "AC Hall effect", though this terminology is in use for other types of Hall effect measurements.

[20] Photons of energy above the semiconductor bandgap are absorbed by interband transitions, as discussed in Chapter 27.

$$E(k_x, k_y, k_z) = \frac{\hbar^2}{2m^\star}(k_x^2 + k_y^2 + k_z^2) \rightarrow (n + \frac{1}{2})\hbar\omega_c + \frac{\hbar^2 k_z^2}{2m^\star}$$

$$\psi = u_{nk}\frac{e^{ik_x x}}{\sqrt{L_x}}\frac{e^{ik_y y}}{\sqrt{L_y}}\frac{e^{ik_z z}}{\sqrt{L_z}} \rightarrow u_{nk}H(x,y)\frac{e^{ik_z z}}{\sqrt{L_z}}, \qquad (25.28)$$

in which two of the degrees of freedom perpendicular to the magnetic field collapse into discrete harmonic oscillator states separated by energy $\hbar\omega_c$ where $\omega_c = qB/m^\star$ is the cyclotron frequency, and $H(x,y)$ is a bound-state wavefunction. The properties related to motion along the magnetic field, if allowed, remain unchanged.

As a result, **the density of states undergoes oscillatory behavior both as a function of energy and the magnetic field**. Therefore, those physical properties of the solid that depend on the density of states, such as the fundamental transport properties, optical properties, and thermodynamic properties such as magnetic susceptibility undergo quantum oscillations as a function of the magnetic field. These quantum oscillations of solids cannot be explained by the semi-classical theories discussed earlier in this chapter.

We shall discuss the quantum magnetic properties in solids starting from the next section. The quantum effect of large magnetic fields on electronic states in solids have many similarities to the effect of a magnetic field on the properties of a single atom. In this section we discuss the effect of magnetic fields on the energies and spectra of atoms. Our goal is to clearly identify and distinguish the effects of a magnetic field on the **orbital motion** and **spin** of electrons, and their energy eigenvalues and eigenfunctions.

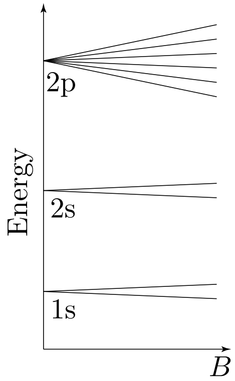


Fig. 25.18 Schematic representation of the magnetic field splitting of energy eigenvalues of an atom. Spin-orbit coupling and very large fields when levels cross are neglected.

The Zeeman effect: The effect of a magnetic field on the optical spectra of an atom was known before the discovery of the electron, nucleus, or quantum theory. The characteristic effect was a **splitting of sharp spectral lines** in frequency $\Delta\nu \propto B$, proportional to the magnetic field as shown in Fig. 25.18. Just before the discovery of the electron, Zeeman observed this effect in hydrogen, and Lorentz gave a classical theory showing that $\Delta\nu = \frac{q}{4\pi m_e}B$, which gave the first experimental evidence of a particle of charge q and mass m_e, about $2000\times$ times smaller than the hydrogen atom. Thus, Zeeman and Lorentz's work deduced the presence of electron and its q/m_e via the effect of magnetic fields on atomic spectra. J. J. Thompson's discovery of the electron confirmed this conclusively shortly after Zeeman's discovery.

Note that the splitting in the frequency has no $\hbar$ in it, and therefore a classical explanation is indeed feasible. As we discussed in Chapter 3, the origin of the spectral lines themselves had remained a mystery till Bohr proposed his orbit theory immediately following Rutherford's discovery of the heavy, positively charged nucleus. We will now see how Bohr's orbit theory shows that $\hbar$ indeed cancels in the Zeeman frequency splitting $\Delta\nu$.

The Bohr magneton: Bohr's hunch that the electron motion did not follow classical mechanics was strengthened by the experimental evidence that several gases and solids responded to external magnetic fields by developing magnetization themselves, but classical mechanics does not allow for any form of magnetism to exist (the Bohr–van Leuween theorem). The essence of Bohr's orbit theory is the **quantization of the angular momentum**, with the classical analogy of an electron spinning around the nucleus in an orbit. This picture is identical to a circular loop carrying a current[21].

For an electron of mass m_e and charge of magnitude q moving in a circular orbit of radius R_n around the nucleus, the Bohr–Sommerfeld quantization condition that gives the correct energy levels of a hydrogen atom in the absence of a magnetic field is $m_e v_n R_n = l\hbar$, where $l = 0, \pm1, \pm2, ...$ is the **orbital quantum number**, a measure of the number of wavelengths that fit the circumference. This implies the effective current flowing in the orbit is $I = q \times (v_n / 2\pi R_n)$, and therefore the magnetic moment of the electron orbiting the nucleus is

$$m_B = IA = q\frac{v_n}{2\pi R_n} \cdot \pi R_n^2 = \frac{1}{2} q v_n R_n = l \times \frac{q\hbar}{2m_e} \implies \boxed{m_B = l\mu_B}$$
$$\text{and the orbital magnetic energy is } E = m_B B \implies \boxed{E = l\mu_B B,} \tag{25.29}$$

where $\mu_B = q\hbar/2m_e$ is called the **Bohr magneton**[22]. In other words, **because of the orbital motion, the electron current behaves as a magnetic dipole of strength an integer multiple of μ_B**. The Bohr magneton is the quantum of magnetic moment, and has a magnitude 0.1 meV/Tesla. As a result, the splitting in energy is $E_{l+1} - E_l = \mu_B B = \frac{q\hbar}{2m_e} B$, which gives a frequency splitting $h\Delta\nu = \frac{q\hbar}{2m_e} B$, implying $\Delta\nu = \frac{q}{4\pi m_e} B$. Note that there is no $\hbar$ in this expression.

Discovery of spin: Fig. 25.19 shows the expected and actually observed splittings of the energy levels of hydrogen in the presence of an external magnetic field. Calling n the principal quantum number, the orbital quantum number l can assume values from 0 to $\pm(n-1)$. The physical meaning of l is shown for the electron clouds in the left of Fig. 25.19. The $l = 0$ states are spherically symmetric, in which the wavefunction may oscillate only in the radial direction. Since it does not have a magnetic moment, its energy level is predicted to be unaffected by a magnetic field in the Bohr orbit theory. For the $n = 1$ 1s ground state, $l = 0$ and therefore the Bohr orbit theory predicts no splitting in the energy level due to a magnetic field. For the $n = 2$ 2s state, $l = 0$ yet again and there should be no magnetic field splitting. Fig. 25.19 left shows that the 2p states have $l = -1, 0, +1$, of which the $l = +1$ and $l = -1$ states are those for which the electron has orbital motion clockwise or counterclockwise around the nucleus. These wavefunctions oscillate both in the radial direction *and simultaneously* around the circular path around the nucleus[23]. Based on the Bohr or-

---

[21] The magnetic moment of a classical loop carrying a current I is $m_B = IA$, where A is the area of the loop. The energy of such a current loop in the presence of an external magnetic field B is $E_+ = +m_B B$, the product of the magnetic moment and the magnetic field if the magnetic moment points along the magnetic field, and $E_- = -m_B B$ if it points in the opposite direction.

[22] Similar to the electron, photon, or phonon, a Bohr magneton is a quantum of magnetic moment.

[23] A crude analogy is an athlete while running around concentric circular tracks in a field, (for no reason!) decides to also cross from the inner to outer lanes simultaneously. The classical analogy of motion for the $l = \pm1$ states is a helical trajectory around the nucleus.

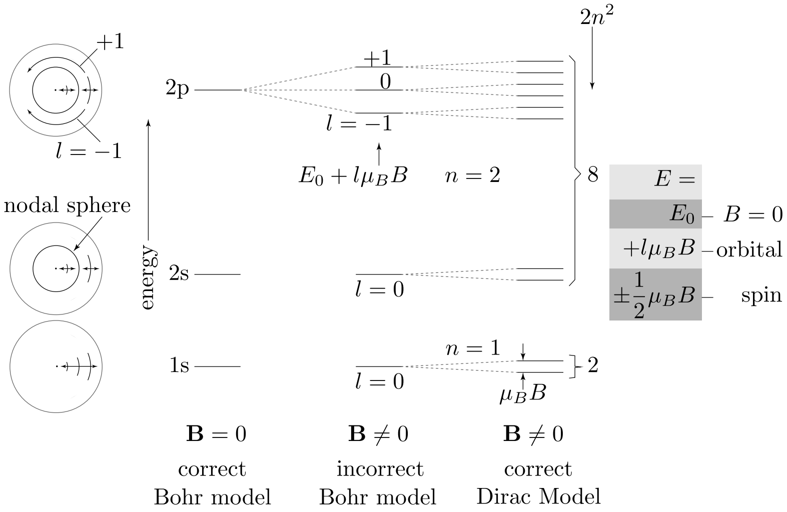


Fig. 25.19 Effect of a magnetic field on the energy levels of a hydrogen atom.

bit theory then, in the presence of a magnetic field, the 2p energy level should split into three levels: $E = E_{2p} - \mu_B B, E_{2p}, E_{2p} + \mu_B B$.

The right side of Fig. 25.19 shows the energy splittings that are actually observed in experiment. Unlike the prediction of the Bohr orbit theory, the $l = 0$ energy 1s eigenvalues are split to energies $E_{1s} \pm \frac{1}{2}\mu_B B$, which is also true for the 2s and 2p $l = 0$ states. The $l = +1$ 2p state is further split to $E_{2p} + \mu_B B \pm \frac{1}{2}\mu_B B$. In other words, each state is split by the magnetic field, in addition to the splitting $l\mu_B B$ due to orbital angular momentum, by an additional $\pm\frac{1}{2}\mu_B B$.

This curious property was explained by Goudsmidt and Uhlenbeck to be due to an **intrinsic angular momentum** of the electron, which does not have any connection to the orbital motion. They called it the **spin** of the electron. The spin is an additional magnetic moment of magnitude $\pm\frac{1}{2}\mu_B$ for electrons. This is why the n^{th} orbital has $2n^2$ states, and not n^2 states as predicted by the Bohr orbit theory. The subsequent introduction of the Pauli spin matrices into the Schrödinger equation, and finally the Dirac equation that merged special relativity with quantum mechanics gave an accurate description of the spin of the electron. Therefore, the end result is the total energy is composed of the energy E_0 determined by the non-magnetic potential, plus $l\mu_B B$ due to orbital angular momentum, and $\pm\frac{1}{2}\mu_B B$ due to the spin angular momentum. The orbital components of the energy thus do have a classical analogy of energies due to magnetic dipoles aligned parallel to, or anti-parallel to the external magnetic field. However, the spin of the electron must be considered to be as fundamental a property as its charge and mass, with or without its orbital angular momentum.

Spin-orbit coupling and fine structure constant: Actually an additional splitting is observed experimentally for the $l \neq 0$ states, beyond the orbital and spin components just discussed. This occurs due to the coupling of spin and orbital motion as we saw in Chapter 12, Section 12.5. Consider the rest frame of the electron. The positively charged nucleus then orbits around the electron. By analogy, electrons in the $l = 0$ electron states see the proton move closer or further, but no orbital motion. The electrons in the $l = 0$ orbitals therefore do not experience any effective magnetic field due to the positive charge of the nucleus. But for $l = \pm 1$ states, or any other $l \neq 0$ states, the positively charged nucleus rotates around the electron in the opposite sense as the electron around the nucleus. The effective current due to the orbital motion therefore creates a bit of effective magnetic field **even in the absence of external magnetic fields**. The energy splitting of the spin-states caused by this orbital motion induced internal magnetic field is called **spin-orbit coupling**. In semiconductor bandstructure, this is the reason the $l \neq 0$ valence band p-states undergo a spin-orbit splitting, whereas the $l = 0$ valence band p-state does not.

This spin-orbit magnetic field seen by the electron due to a spinning positive charge at radius R is $B_{SO} = \frac{\mu_0 I}{2R}$, where $\mu_0 = \frac{1}{c^2\epsilon_0}$ is the permeability of vacuum, c is the speed of light, ϵ_0 is the permittivity of vacuum, and $I = qv/(2\pi R)$ is the current. Then using the Bohr–Sommerfeld quantization condition, for the 2p $l = +1$ state, we have $B_{SO} = \frac{\mu_0 q\hbar}{4\pi m_e R^3} = \frac{q}{4\pi\epsilon_0} \cdot \frac{\hbar}{m_e c^2 R^3}$. Therefore, the energy splitting due to spin-orbit coupling of the electron energy is $E_{SO} = \pm\frac{1}{2}\mu_B B_{SO} = \pm\frac{q^2}{4\pi\epsilon_0} \cdot \frac{\hbar^2}{4m_e^2 c^2 R^3}$. The ratio of the spin-orbit energy at $R = a_B^0$, the Bohr radius, and the ground state energy of the hydrogen atom is then $E_{SO}/E_{1s} = \alpha^2$ where $\alpha = \frac{q^2}{4\pi\epsilon_0} \cdot \frac{1}{\hbar c} = \frac{1}{137}$ is the **fine structure constant**. Therefore, the spin-orbit splitting for the hydrogen atom 2p states is $\alpha^2 \times 13.6$ eV ≈ 0.72 meV. Though small, it is experimentally detectable. As was discussed in Chapter 12, the spin-orbit coupling can be strongly enhanced in semiconductors with heavy nuclei since the number of protons, and hence the charge and current I increase: this is why the spin-orbit coupling energy in narrow bandgap semiconductors can be large, sometimes even larger than the energy bandgap! It is important to remember spin-orbit splitting occurs even when $\mathbf{B} = 0$.

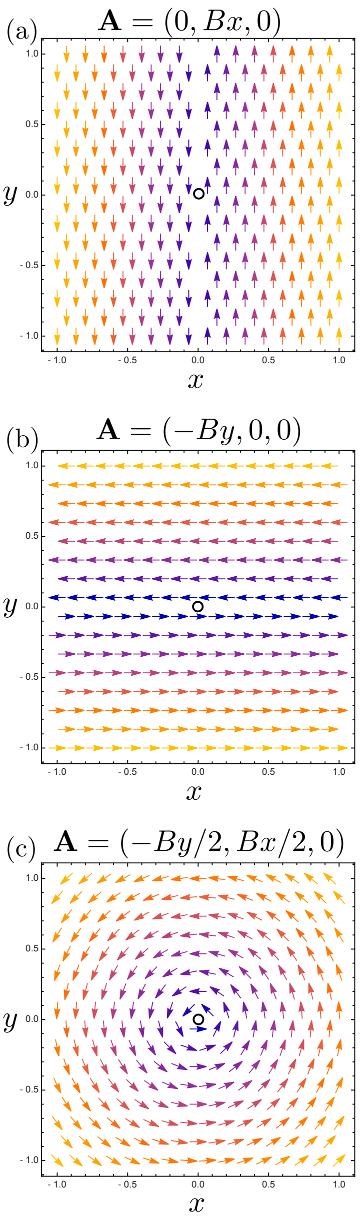


Fig. 25.20 Three gauge choices for the vector potential **A** that lead to the same z-directed magnetic field $\mathbf{B} = B\hat{z}$. A charged particle experiences a rotational force around the center due to the magnetic field, via the vector potential.

25.8 Landau levels in solids

The effect of a magnetic field on the properties of electrons in any scenario is evaluated by introducing into the Schrödinger equation the effective magnetic "potential energy" experienced by electrons. This will be our passport to understanding quantum magnetotransport and magnetooptical properties in solids. The first step to achieve that was taken by Lev Landau (Fig. 25.4), who solved the problem for free–electrons in 3D as a model for conduction electrons in a metal. Using Landau's method, we will follow the same general strategy as in Chapter 5 to first evaluate the eigenvalues and eigenfunctions of electrons in a solid in a magnetic field. From the eigenvalues, we will obtain the density of states, which, with the Fermi occupation function, will play the central role in understanding quantum magnetotransport oscillations: both the Shubnikov de Haas, and the quantum Hall effects.

Since the Schrödinger equation requires potentials and not forces, a magnetic field $\mathbf{B}$ enters it via its vector potential $\mathbf{A}$, such that $\mathbf{A} = \nabla \times \mathbf{A}$. A rigorous justification is obtained from the Hamiltonian in classical mechanics, where for a charge Q the total momentum in a magnetic field is $\mathbf{p} - Q\mathbf{A}$. $Q\mathbf{A}$ has units of momentum. $\hat{\mathbf{p}} \rightarrow \hat{\mathbf{p}} - Q\mathbf{A}$ is the **Peierls–Berry substitution** in quantum mechanics.

For the rest of the chapter, a magnetic field $\mathbf{B} = B\hat{z}$ is discussed. Several **gauge choices** of $\mathbf{A}$ can lead to this same magnetic field. Three popular $\mathbf{A}$ are $(0, Bx, 0)$, $(-By, 0, 0)$, and $(-By/2, Bx/2, 0)$, as is easily verified by taking the curls. Fig. 25.20 shows these three gauges. Thinking of $\mathbf{A}$ as a momentum, the vectors give the sense of a counterclockwise torque on a particle at the origin in each of the three cases.

We choose the first gauge, and postpone spin for later. The quantum solutions for electrons (charge $Q = -q$) in a solid with an isotropic effective mass $m^\star$ are

$$\boxed{\frac{(\hat{\mathbf{p}}+q\mathbf{A})^2}{2m^\star}\psi = E\psi.}\ \text{For } \mathbf{A} = \begin{bmatrix}0\\ Bx\\ 0\end{bmatrix}, \frac{1}{2m^\star}\left(\begin{bmatrix}\hat{p}_x\\ \hat{p}_y\\ \hat{p}_z\end{bmatrix} + q\begin{bmatrix}0\\ Bx\\ 0\end{bmatrix}\right)^2\psi = E\psi$$

$$\psi = e^{ik_y y}e^{ik_z z}\phi(x) \text{ gives } \left[\frac{\hat{p}_x^2}{2m^\star} + \frac{m^\star\omega_c^2}{2}\left(x + \frac{\hbar k_y}{qB}\right)^2\right]\phi = \left(E - \frac{\hbar^2 k_z^2}{2m^\star}\right)\phi$$

$$\Longrightarrow \boxed{E = \left(n+\frac{1}{2}\right)\hbar\omega_c + \frac{\hbar^2 k_z^2}{2m^\star}} \text{ and } \boxed{\psi_{k_y,k_z} = \frac{e^{ik_z z}e^{ik_y y}}{\sqrt{L_z L_y}}H_n\left(x + \frac{\hbar k_y}{qB}\right).} \quad (25.30)$$

We recognize the equation as that of a **harmonic oscillator** in x, but with the **origin shifted to** $x_0 = -(\hbar k_y)/(qB)$. The plane wave solution along z is expected, but the one along y is not: it is a consequence of the gauge choice, and will be seen now to not lead to any *experimentally accessible* feature that prefers the y-direction over the x-direction. The $(-By/2, Bx/2, 0)$ gauge can make this explicit in cylindrical coordinates, but the gauge here lets us work in the simpler Cartesian coordinates (see Exercise 25.6).

Consider the 3D Fermi sphere in k-space in Fig. 25.21 (a). The energy eigenvalues in Equation 25.30 do not depend on (k_x, k_y) explicitly, but on their combined quantum number $n = 0, 1, 2, ...$ But the eigenvalues depend continuously on k_z. Thus, all allowed energy eigenvalues in k-space must collapse onto the surfaces of the concentric **Landau cylinders** shown, each for a distinct n. The corresponding energy bandstructure is shown in Fig. 25.21 (b): the energies are separated exactly by the cyclotron energy $\hbar\omega_c$, with the minimum at $\frac{1}{2}\hbar\omega_c$. The corresponding density of states collapses from the 3D $\sqrt{E}$ form into peaks called **Landau levels** as shown in Fig. 25.21 (c). This looks like the 1D DOS with several subbands. Let's see why.

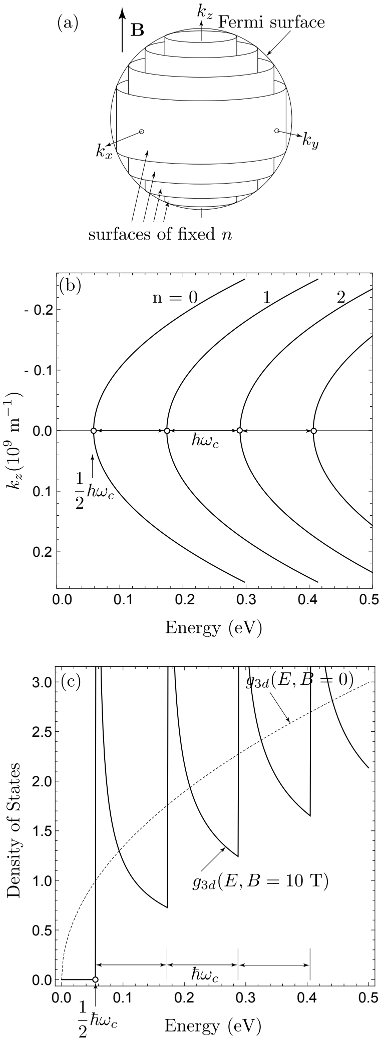


Fig. 25.21 Density of states of 3D electron system modified by a magnetic field showing the formation of Landau levels. The density of states is calculated for $m^\star = 0.01 m_e$ and $B = 10$ T, and the DOS is in the units of 10^{43}/J.m^3 for which the Landau level separation is $\hbar\omega_c = \frac{\hbar q B}{m^\star} = 0.116$ eV. Since a magnetic field along z enforces cyclotron orbits in the $x - y$ direction, it "freezes-out" two degrees of freedom. Therefore, the 3D DOS (dashed line) becomes a series of 1D DOS (solid lines), with free 1D motion along the direction of the magnetic field still allowed.

In a sample of size $L_x \times L_y \times L_z$, the harmonic oscillator wavefunction solution[24] extends over a characteristic length scale $l_B = \sqrt{\hbar/qB}$, called the **magnetic length**. It is independent of the electron effective mass and depends only on the magnetic field, taking a value $l_B \approx 26$ nm at $B = 1$ Tesla. The eigenfunctions in Equation 25.30 thus are plane waves along the z- and y- directions, but are localized in the x-direction to a length $\sim l_B$. A large number of such "strips" of area $l_B L_y$ fit in the solid (Fig. 25.22). Exactly how many determines the degeneracy of a Landau level corresponding to each n. This degeneracy is obtained by requiring that the center of the x-portion stays within the solid. Since $k_y = n_y \times (2\pi/L_y)$, the degeneracy is:

$$0 \leq \frac{\hbar k_y}{qB} \leq L_x \implies n_y \frac{2\pi}{L_y} \leq \frac{qBL_x}{\hbar} \implies \boxed{n_y \leq \frac{BL_xL_y}{h/q} = \frac{L_xL_y}{2\pi l_B^2} = \frac{\phi}{\phi_0}}, \tag{25.31}$$

where we recognize $\phi_0 = h/q$ as the standard (not superconducting!)[25] **magnetic flux quantum**, and $\phi = BL_xL_y = \oint \mathbf{B} \cdot d\mathbf{S}$ as the total magnetic flux through the sample. Thus, the **degeneracy of a Landau level is equal to the number of magnetic flux quanta passing through the sample** at a magnetic field B, as illustrated in Fig. 25.22.

Consider Fig. 25.21 (b): each subband shown in a ordinary quasi-1D conductor would have only a spin degeneracy of 2. But each k_z-state of each subband here has a massively large degeneracy: equal to the value of Equation 25.31. This is because all the (k_x, k_y) states of the 3D bandstructure within an energy window of $\hbar\omega_c$ have collapsed into the closest 1D subband. Therefore, we can write the density of states (per unit 3D volume) in a straightforward manner:

$$g(E,B) = \frac{\phi/\phi_0}{L_xL_y} \cdot g_{1d}(E) = \frac{1}{2\pi l_B^2} \cdot \frac{g_s}{2\pi}\sqrt{\frac{2m^\star}{\hbar^2}} \sum_{n=0,1,2,...} \frac{1}{\sqrt{E - (n+\frac{1}{2})\hbar\omega_c}}, \tag{25.32}$$

which is plotted in Fig. 25.21 (c) and compared to the corresponding 3D DOS at $\mathbf{B} = 0$. Note the **oscillatory** nature of the large $\mathbf{B} \neq 0$ DOS, each of which peak is a Landau level. This oscillatory feature is responsible for quantum oscillations in transport and thermodynamic properties related to the DOS. Before we proceed to discuss these properties, we do three extensions to this model to prepare for comparison with experimental data: (a) include spin, (b) obtain the corresponding results for 2D systems, and (c) account for disorder.

Spin: The inclusion of spin via Pauli spin matrices changes the Hamiltonian and the corresponding energy eigenvalues to[26]

$$[\frac{(\hat{\mathbf{p}} + q\mathbf{A})^2}{2m^\star} + g\mu_B\hat{\sigma}\cdot\mathbf{B}]\psi = E\psi \implies E = (n + \frac{1}{2} \pm \underbrace{\frac{gm^\star}{4m_e}}_{=\gamma/2})\hbar\omega_c + \frac{\hbar^2k_z^2}{2m^\star} \tag{25.33}$$

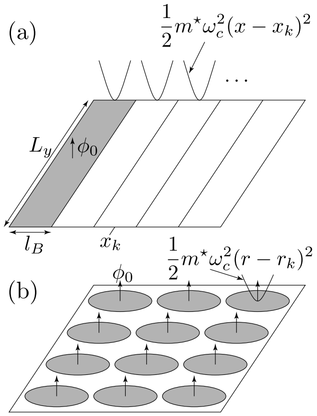

(a)
$\frac{1}{2}m^\star\omega_c^2(x-x_k)^2$
...
L_y
ϕ_0
l_B
x_k
(b)
ϕ_0
$\frac{1}{2}m^\star\omega_c^2(r-r_k)^2$

Fig. 25.22 (a) Illustration that each Landau level corresponds to a magnetic flux quantum for the gauge $\mathbf{A} = B(0,x,0)$. (b) The same as (a) but for the circular gauge $\mathbf{A} = \frac{B}{2}(-y,x,0)$. It is as if the magnetic field has created ϕ/ϕ_0 "atoms" in the solid, whose density changes with B.

[24] The harmonic oscillator solution is $H_n(x) = \frac{1}{\sqrt{2^n n!}}(\frac{1}{\pi l_B^2})^{\frac{1}{4}} e^{-\frac{(x-x_0)^2}{2l_B^2}} \mathcal{H}_n(\frac{x-x_0}{l_B})$, where $\mathcal{H}_n(u)$ are Hermite polynomials.

[25] In superconductors, the flux quantum is $\phi_0 = h/2q$ because the charge carriers are not single electrons, but Cooper pairs of two electrons.

[26] Here g is the Lande g factor, which is equal to $g = 2$ for the free–electron, but in direct bandap semiconductors for conduction band states it is $g = 2(1 - \frac{\frac{m_e}{m_c^\star}-1}{2+3\frac{E_g}{\Delta}})$, where $m_c^\star$ is the conduction band effective mass, E_g the bandgap, and Δ the spin-orbit splitting energy. For example, in InAs $m_c^\star = 0.023m_e$, $\Delta = 0.41$ eV, and $E_g = 0.354$ eV gives $g = -16.5$, close to the experimental value of -15. The g-factor for indirect bandgap semiconductors is $g \approx 2$.

and the corresponding DOS is

$$g(E,B) = \frac{1}{(2\pi l_B)^2}\sqrt{\frac{2m^\star}{\hbar^2}}\Big[\sum_{n=0,1,2,...}\frac{1}{\sqrt{E-(n+\frac{1}{2}+\frac{\gamma}{2})\hbar\omega_c}} + \sum_{n=0,1,2,...}\frac{1}{\sqrt{E-(n+\frac{1}{2}-\frac{\gamma}{2})\hbar\omega_c}}\Big]. \tag{25.34}$$

2D Electron systems: We can imagine that the magnetic field forces cyclotron orbits of electrons in a plane perpendicular to it. Thus, it effectively localizes, or "freezes-out" two degrees of free motion perpendicular to it. This is the reason the 3D DOS collapses into a series of 1D Landau levels, whereas a 2D DOS collapses into a series of atomic-like discrete, or 0D Landau levels. Consider a 2D system such that electrons (or holes) can only move in the (x, y) plane, perpendicular to $\mathbf{B} = B\hat{z}$. The solutions may be written down directly from the corresponding 3D case with no k_z:

$$\Big[\frac{(\hat{\mathbf{p}}+q\mathbf{A})^2}{2m^\star} + g\mu_B\hat{\sigma}\cdot\mathbf{B}\Big]\psi = E\psi \implies E = (n+\frac{1}{2}\pm\frac{\gamma}{2})\hbar\omega_c$$
$$g(E,B) = \frac{1}{l_B^2}\Big(\sum_{n=0,1,2,...}\delta[E-(n+\frac{1}{2}+\frac{\gamma}{2})\hbar\omega_c] + \sum_{n=0,1,2,...}\delta[E-(n+\frac{1}{2}-\frac{\gamma}{2})\hbar\omega_c]\Big), \tag{25.35}$$

which is, not surprisingly, a series of Dirac delta functions at the Landau level energies.

Disorder: The final ingredient needed to explain experimental data is the effect of disorder[27]. Let's find the probability of occupation of a quantum state $E_n = \hbar\omega_n$ due to scattering to other states $E = \hbar\omega$ within a quantum scattering time τ. Obtaining the probability of transitions from the time-dependent Schrödinger equation (Chapter 20), and averaging it over the quantum scattering time τ gives

$$\frac{1}{2}\Big|\int_0^t dt' e^{i(\omega-\omega_n)t'}\Big|^2 = \frac{1-\cos(\omega-\omega_n)t}{(\omega-\omega_n)^2} \implies$$
$$\frac{\int_0^\infty dt\frac{1-\cos(\omega-\omega_n)t}{(\omega-\omega_n)^2}e^{-\frac{t}{\tau}}}{\int_0^\infty dt e^{-\frac{t}{\tau}}} = \frac{1}{(\omega-\omega_n)^2+\frac{1}{\tau^2}}, \quad \boxed{P(E) = \frac{\hbar/(2\pi\tau)}{(E-E_n)^2+(\frac{\hbar}{\tau})^2}}. \tag{25.36}$$

If in the absence of disorder the energy is E_n, then the probability of an electron possessing an energy E is given[28] by the boxed Equation 25.36. Fig. 25.23 (a) and (b) show the corresponding DOS with disorder

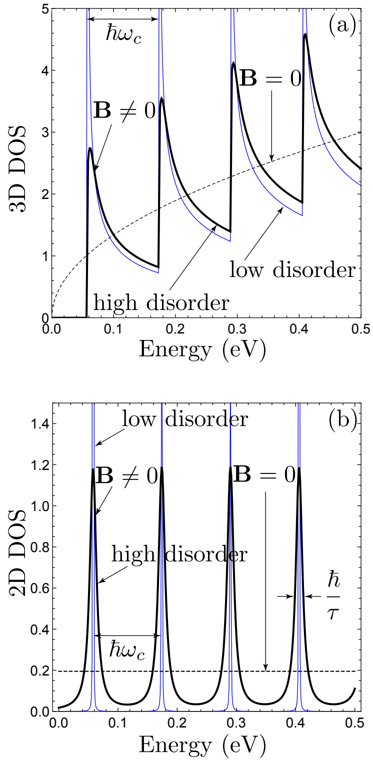


Fig. 25.23 (a) 3D and (b) 2D density of states in the presence of a large magnetic field showing Landau levels. The constant 2D DOS collapses into Landau levels that mimic a series of delta functions with gaps equal to $\hbar\omega_c$. Due to disorder, the DOS broadens as indicated schematically.

[27] The effect of disorder is modeled in a simple way by changing the energy $E \to E + i\Gamma$ in the DOS of equations 25.34 and 25.35, and taking the real part of the expressions. The energy Γ characterizing disorder is $\Gamma = \hbar/\tau$, where τ is a quantum scattering rate of the electron states due to disorder.

[28] Note that $\int dE \cdot P(E) = 1$. This is the origin of Lorentzian broadening.

induced broadening for 3D and 2D carriers. The dashed lines are the DOS in the absence of a magnetic field. The magnetic field creates the Landau level oscillations in the DOS as shown by thin lines. The effect of disorder is to wash out the sharp oscillations in the DOS as shown by thick lines. It is seen in Fig. 25.23 (b) that the criteria for clear oscillations is $\hbar\omega_c > \hbar/\tau$, or $\omega_c\tau > 1$. As a result of disorder all Dirac delta functions of energy are replaced by $\delta(E - E_n) \to \frac{\Gamma/2\pi}{(E-E_n)^2+\Gamma^2}$ where $\Gamma = \hbar/\tau$, and the amplitude of quantum magnetoresistance oscillations is damped by a factor $\exp\left[-\pi/(\omega_c\tau)\right]$ due to scattering. Let us now proceed to discuss experimental data and use the methods discussed in this section to understand them.

Fig. 25.24 L. Shubnikov (left), de Haas (middle), and van Alphen (right). In 1930 Shubnikov and de Haas reported the observation of oscillations of magnetoresistance of crystalline Bismuth, shortly followed by a report by de Haas and van Alphen of oscillations of magnetic susceptibility in the same material.

25.9 Shubnikov de Haas effect

In 1930, Shubnikov and de Haas (Fig. 25.24) observed that at low temperatures, the **longitudinal resistivity** ρ_{xx} of Bismuth semimetal of very high purity exhibited oscillations as a function of the magnetic field. For the same material and under similar conditions, de Haas and van Alphen also reported oscillations in the **magnetic susceptibility**. Though Landau had, around the same time, solved the quantum mechanical problem of free–electrons in a magnetic field and explained the diamagnetism of metals, he had ruled out the possibility of oscillations in susceptibility as "unobservable". Thus the reports of susceptibility and magnetoresistance oscillations were remarkable.

Shubnikov de Haas oscillations in magnetoresistance were observed in the high electron mobility semiconductors InSb and InAs in late 1950s, and interestingly in the surface inversion 2DEGs in silicon MOSFET structures in 1966. Since then such oscillations have been observed in most semiconductor structures that meet the criteria $\omega_c\tau > 1$. The de Haas van Alphen susceptibility oscillations were explained early by an extension of Landau's theory by Rudy Peierls (Fig. 8.5) within a few years of its observation in 1930. The theory for Shubnikov de Haas magnetoresistance oscillations had to wait for nearly 30 years, till late 1950s. This is because of added complexity: it represents a case where the Boltzmann transport equation is inadequate and a density matrix approach is needed, as discussed later.

Fig. 25.25 (a) shows a qualitative explanation for magnetoresistance oscillations, which we shall use as a guide for quantitative understanding. Electrons in the semiconductor fill energy states till the Fermi level E_F which in the figure lies in the middle of two Landau-level peaks. The longitudinal electronic conductivity is proportional to the density of states at the Fermi level, $\sigma \propto g(E_F)$. Consider first the case where a large magnetic field is kept constant, meaning the DOS is fixed in Fig. 25.25 (a). Suppose now the free carrier concentration is changed by some means, changing E_F. The conductivity $\sigma(E_F)$ will then show Shubnikov de Haas oscillations at a fixed magnetic field as E_F sweeps through the DOS. This case is achieved for 2D carriers by field-effect

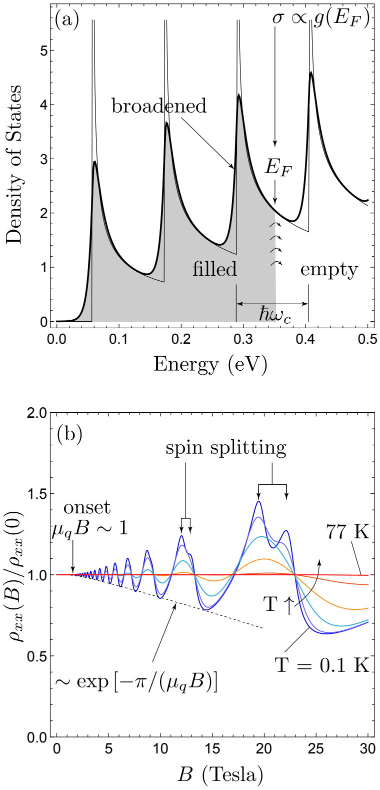


Fig. 25.25 Shubnikov de Haas oscillations in the longitudinal magnetoresistance calculated for an effective mass $m_c^\star = 0.2m_e$, $g = 2$, quantum scattering time $\tau_q = 10^{-12}$ s, and 3D carrier concentration $n_{3d} = 10^{18}$ /cm^3. Note the envelope of the oscillations takes the form $\sim \exp\left[-\pi/(\mu_q B)\right]$.

with a gate: then oscillations are observed as a function of the gate voltage. This is the path to the quantum Hall effect to be discussed in the next section. Consider next the situation when the magnetic field is increased for a fixed free carrier concentration. Then, $\hbar\omega_c = \hbar qB/m^\star$ increases and thus the Landau level spacing increases, and σ initially increases and then decreases if E_F does not change appreciably (to be exact, it does change in 3D!). Thus, with changing B the conductivity (and resistivity) must oscillate as seen in Fig. 25.25 (b).

Fig. 25.25 (b) shows calculated Shubnikov de Haas oscillations at several temperatures. They exhibit five characteristic features:

(a) the onset at $\mu_q B \sim 1$ where $\mu_q = q\tau_q/m^\star$ is the quantum mobility,
(b) oscillations periodic in E_F/B,
(c) thermal damping as $\frac{\chi}{\sinh\chi}$ where $\chi = 2\pi^2(k_bT)/(\hbar\omega_c)$,
(d) disorder-induced damping as $\exp[-\pi/(\mu_q B)]$, and
(e) spin splitting at the highest B fields.

We now outline a theoretical model for the calculation and analysis of SdH oscillations. Let's begin from a rather interesting historical note on **why the magnetic susceptibility of free–electrons is zero in classical physics** – which in itself is at odds with experiment, and how quantum mechanics resolves this puzzle. Consider Fig. 25.26 which shows the classical response of "free" charges confined in a 2D plane to a magnetic field B perpendicular to the plane. The charges move in cyclotron orbits of radius $R = mv/(qB)$ and current $I = qv/(2\pi R)$. Far from the edges they complete their orbits and each produce a magnetization $m = I \cdot \pi R^2 = (\frac{1}{2}mv^2)/B$. Therefore the net magnetization for N bulk charges is $Nm = (\frac{N}{2}mv^2)/B$. But this is clearly absurd because the magnetization becomes infinite at zero B!

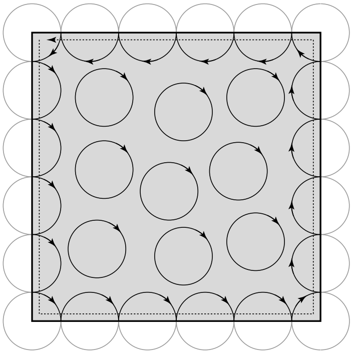


Fig. 25.26 Lorentz's argument for zero diamagnetism. The closed loops inside the sample have a net magnetic moment pointing into the paper, and proportional to the sum of the areas of the circles. The "skipping" orbits at the edge have far less orbits, but enclose a much larger net area that produces a magnetization pointing out of the paper, opposite to the closed loops. It is provable that they cancel exactly! The edge currents in the quantum Hall effect will be similar to these semi-classical edge currents.

Hendrik Lorentz's (Fig. 26.6) resolution of this conundrum consists the following argument. The orbits at the edge of the sample "skip" along and complete a loop around the perimeter. Note that such edge-crawling electrons produce an **edge current** that flows in a direction *opposite* to all the bulk orbits. The edge current is the same as each of the bulk loop currents, but the area of the edge path is much larger (= entire sample!). Since the number of charges at the edge is much smaller than the interior, the increased area balances out exactly, and the net magnetization from the classical orbits is zero.

The fact that net classical magnetization is zero is proven in a simpler manner from a thermodynamic approach. The magnetization is $M = -\partial F/\partial B$, where $F = U - TS$ is the free energy at temperature T, U is the total energy, and $S = -k_b \ln Z$ is the entropy, where $Z = \sum \exp[-E_n/k_bT]$ is the classical partition function. In classical mechanics, the energy is simply $E = (\mathbf{p} + q\mathbf{A})^2/2m^\star$ where $\mathbf{A}$ is the magnetic vector potential. In the calculation of the partition function the sum over momenta converts into an integral of a Gaussian of the form $\exp[-(\mathbf{p} + q\mathbf{A})^2/2m^\star k_bT]$. It is clearly seen[29] then that the value of the integral is independent of $\mathbf{A}$, and therefore $M = -\partial F/\partial B = 0$. This is the **Bohr–van Leeuwen theorem** stating that there can be no net magnetization for free–electrons in classical mechanics.

[29] The equivalent mathematical result is $\int_{-\infty}^{+\infty} du \cdot e^{-u^2} = \int_{-\infty}^{+\infty} du \cdot e^{-(u-A)^2}$: the total area under a Gaussian is the same no matter where the origin.

But experimentally some metals exhibit weak paramagnetism ($\chi_m = \partial M/\partial B > 0$), while others exhibit diamagnetism ($\chi_m = \partial M/\partial B < 0$). The paramagnetism of metals is due to electron spin and the Pauli exclusion principle, as was first explained by Pauli without considering orbital effects or Landau quantization. Writing the non-quantized free-electron model 3D density of states as $g_0(E) = \frac{1}{2\pi^2}(\frac{2m^\star}{\hbar^2})^{3/2}\sqrt{E}$, Pauli found a **positive susceptibility** χ_m simply by including in the free energy[30] the spin of electrons of volume density n in a metal:

$$F_0(E_F) = U - TS = nE_F - k_bT\int_0^\infty dE g_0(E)\ln\left[1+e^{\frac{E_F-E}{k_bT}}\right]$$

$$= nE_F - (k_bT)^2 g_0(k_bT)\cdot\frac{\sqrt{\pi}}{2}\mathcal{F}_{\frac{3}{2}}\left(\frac{E_F}{k_bT}\right)$$

From Fig. 25.27, $F_{\text{spin}} = \frac{1}{2}[F_0(E_F+\mu_B B) + F_0(E_F-\mu_B B)]$

$$\Longrightarrow \boxed{\chi_m(\text{para.}) = -\frac{\partial^2 F_{spin}}{\partial B^2} = +\mu_B^2 g_0(E_F). \text{ Pauli paramagnetism}} \tag{25.37}$$

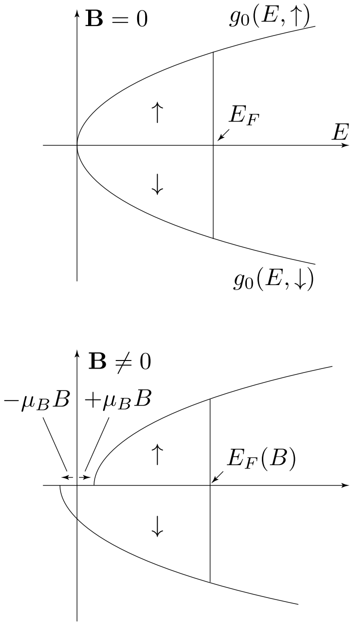


Fig. 25.27 Illustration of Pauli paramagnetism due to electron spin. In a magnetic field B, the energy eigenvalues of up and down spin electrons change in opposite directions by $\pm\mu_B B$ where $\mu_B = q\hbar/2m_e$ is the Bohr magneton. At thermal equilibrium, a single Fermi level implies larger spins aligned with the field, and a positive magnetization and positive susceptibility.

On the other hand, **diamagnetism** is due to the formation of Landau levels by orbital quantization. Understanding it needs more work, but it holds the secrets of the SdH and dHvA effects, as we now discuss. Using the 3D energy eigenvalues obtained for electrons in a magnetic field in Equation 25.33, the free energy and magnetization are obtained for a 3D Fermi gas of electrons in the following steps:

$$E_F \text{ is defined by total electrons } N = \sum_{k_z,n,\nu}\frac{1}{1+e^{\frac{(n+\frac{1}{2}\pm\frac{\gamma}{2})\hbar\omega_c+\frac{\hbar^2k_z^2}{2m^\star}-E_F}{k_bT}}}.$$

$$\text{The free energy is } F = U - TS = NE_F - k_bT\sum_{n,k_z,s}\ln\left[1+e^{\frac{E_F-E_{n,k_x,s}}{k_bT}}\right]$$

$$= NE_F - k_bT\int_0^\infty dE\cdot g(E,B)\cdot\ln\left[1+e^{\frac{E_F-E}{k_bT}}\right] = NE_F -$$

$$k_bT\int_0^\infty dE\cdot\underbrace{\frac{1}{(2\pi l_B)^2}\sqrt{\frac{2m^\star}{\hbar^2}}\sum_{n=0}^{\infty}\frac{1}{\sqrt{E-(n+\frac{1}{2}\pm\frac{\gamma}{2})\hbar\omega_c}}}_{g(E,B)}\cdot\underbrace{\ln\left[1+e^{\frac{E_F-E}{k_bT}}\right]}_{H(E)}, \tag{25.38}$$

where the quantized density of states $g(E,B)$ due to the magnetic field is used from Equation 25.34 (see Fig. 25.28). We define single and double integrals of the density of states $g(E)$ with $g(0) = 0$ to obtain

$$g_1(E) = \int_0^E du\cdot g(u), \text{ and } g_2(E) = \int_0^E du\cdot g_1(u) = \int_0^E dv\int_0^v du\cdot g(u),$$

$$\Longrightarrow \int_0^\infty dE\cdot g(E)\cdot H(E) = \int_0^\infty dE\left(-\frac{\partial H}{\partial E}\right)g_1(E) = \int_0^\infty dE\left(\frac{\partial^2 H}{\partial E^2}\right)g_2(E). \tag{25.39}$$

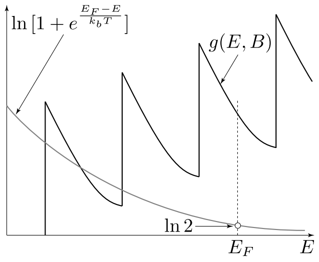


Fig. 25.28 Illustration of terms in Landau free energy.

---

[30] The magnetization M is obtained from the free energy F using $M = -\partial F/\partial B$. The magnetic susceptibility is $\chi_m = -\partial^2 F/\partial B^2$. See Exercise 25.8. Here $\mathcal{F}_j(\eta)$ is the Fermi–Dirac integral of order j.

The identities above (as can be verified by successive partial integration) hold for a function $H(E)$ with the property $H(E \to +\infty) \to 0$. We recognize that in the last line of Equation 25.38: $H(E) = \ln[1 + \exp[(E_F - E)/k_bT]]$ has the property $\partial H/\partial E = -f_0/k_bT$, and $\partial^2 H/\partial E^2 = (-\partial f_0/\partial E)/k_bT$ where f_0 is the Fermi–Dirac function. Using Equation 25.39 and using a spin degeneracy of $g_s = 2$ for the moment (we shall reinstate the spin splitting later!), we have

$$\begin{aligned} g_2(E) &= \frac{2}{(2\pi l_B)^2}\sqrt{\frac{2m^\star}{\hbar^2}}\frac{4}{3}(E-(n+\frac{1}{2})\hbar\omega_c)^{\frac{3}{2}} \\ F = NE_F - \int_0^\infty dE(-\frac{\partial f_0}{\partial E})g_2(E) &= \\ NE_F - \int_0^\infty dE(-\frac{\partial f_0}{\partial E})\sum_{n=0}^{\infty}\frac{2}{(2\pi l_B)^2}\sqrt{\frac{2m^\star}{\hbar^2}}\frac{4}{3}&(E-(n+\frac{1}{2})\hbar\omega_c)^{\frac{3}{2}}. \end{aligned} \tag{25.40}$$

If the last integral is approximated by first replacing the summation over the Landau levels n by an integral over Δn, it results in the incorrect result of zero dependence on magnetic field. Therefore the summation must be evaluated exactly. A remarkable result called the **Poisson summation formula** comes to the rescue, which we state without proof (see Fig. 25.29 for an illustration of this formula):

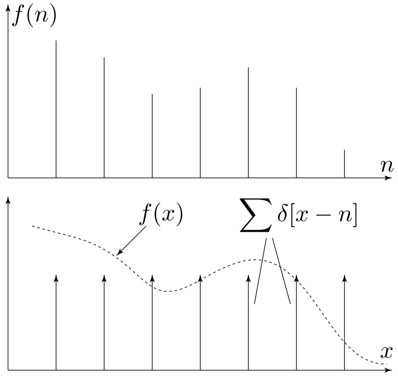


Fig. 25.29 Illustration of Poisson summation formula. The sum of the discrete series $\sum_n f(n)$ is converted to an exact integral by multiplying $f(x)$ with a Dirac delta comb function and using its Fourier representation $\sum_{n=-\infty}^{+\infty}\delta[x-n] = \sum_{r=-\infty}^{+\infty} e^{2\pi i r x}$.

$$\begin{gathered} \sum_{n=-\infty}^{+\infty}\delta[x-n] = \sum_{r=-\infty}^{+\infty} e^{2\pi i r x} \implies \sum_{n=-\infty}^{+\infty} f(n) = \sum_{r=-\infty}^{+\infty}\int_{-\infty}^{+\infty} dx f(x) e^{2\pi i r x} \\ \sum_{n=0}^{\infty} f(n+\frac{1}{2}) = \int_0^\infty dx f(x) + \frac{f'(0)}{24} - \sum_{r=1}^{\infty}\frac{(-1)^r}{2\pi^2 r^2}\int_0^\infty dx f''(x)\cos(2\pi r x) \\ \text{and the Fresnel integral is: } \int_0^\infty dx\cos(a \pm bx^2) = \frac{1}{2}\sqrt{\frac{\pi}{b}}\cos(a \pm \frac{\pi}{4}). \end{gathered} \tag{25.41}$$

Using the Poisson summation formula and the Fresnel integral in the limit $E >> \hbar\omega_c$, the sum over Landau levels evaluates to

$$\begin{aligned} \sum_{n=0}^{\infty}(E-(n+\frac{1}{2})\hbar\omega_c)^{\frac{3}{2}} &= \\ (\hbar\omega_c)^{\frac{3}{2}}\Big[\frac{2}{5}(\frac{E}{\hbar\omega_c})^{\frac{5}{2}} - \frac{1}{16}(\frac{E}{\hbar\omega_c})^{\frac{1}{2}} - \frac{3}{8\sqrt{2}\pi^2}\sum_{r=1}^{\infty}\frac{(-1)^r}{r^{\frac{5}{2}}}&\cos(2\pi r\frac{E}{\hbar\omega_c} - \frac{\pi}{4})\Big]. \end{aligned} \tag{25.42}$$

Substituting this form in Equation 25.40, the integral for the free energy F now can be evaluated exactly with the help of two integrals. In Chapter 4, Exercise 4.12 we introduced integrals of the Fermi–Dirac function with the promise that they would be needed for magnetotransport. It is time to keep that promise, and the integrals are reproduced here. For $\beta = 1/(k_bT)$ and $f_0(E) = 1/(\exp[\beta(E - E_F)] + 1)$ the

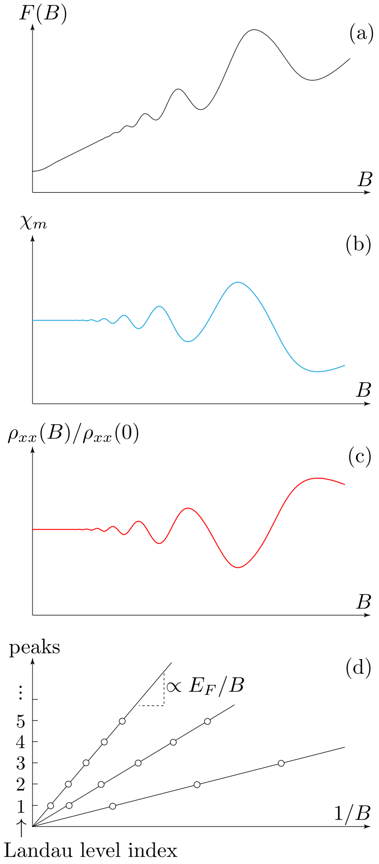


Fig. 25.30 Illustration of (a) oscillation of free energy with magnetic field, and the corresponding oscillations in the (b) magnetic susceptibility (de Haas van Alphen effect), and (c) the longitudinal magnetoresistance (the Shubnikov de Haas effect). (d) Since the oscillations are periodic in E_F/B, if the peaks locations are plotted vs. $1/B$, they form a **Landau fan diagram** as shown, with the corresponding Landau levels on the left axis. The slope is proportional to E_F which depends on the carrier concentration.

Fermi–Dirac function,

$$\int_{-\infty}^{+\infty} dE(-\frac{\partial f_0}{\partial E})e^{ikE} = \frac{\frac{\pi k}{\beta}}{\sinh{(\frac{\pi k}{\beta})}}, \text{ and}$$

$$\int_{-\infty}^{+\infty} dE(-\frac{\partial f_0}{\partial E})\cos{(2\pi s\frac{E}{\hbar\omega_c} - \frac{\pi}{4})} = \frac{2\pi s^2 \frac{k_b T}{\hbar\omega_c}}{\sinh{(2\pi s^2 \frac{k_b T}{\hbar\omega_c})}}\cos{(2\pi s\frac{E_F}{k_b T} - \frac{\pi}{4})}. \tag{25.43}$$

Combining all of the above, the free energy of electrons in a quantizing magnetic field is written in terms of the density of states in the absence of magnetic fields $g_0(E) = \frac{1}{2\pi^2}(\frac{2m^\star}{\hbar^2})^{3/2}\sqrt{E}$ as:

$$F = F(B=0) + \frac{1}{24}(\hbar\omega_c)^2 g_0(E_F) + \frac{1}{(2\pi)^2}(\hbar\omega_c)^2 g_0(E_F)\sqrt{\frac{\hbar\omega_c}{2E_F}} \times \sum_{r=1}^{\infty}\frac{(-1)^r}{r^{5/2}}\frac{2\pi^2 r\frac{k_b T}{\hbar\omega_c}}{\sinh{(2\pi^2 r\frac{k_b T}{\hbar\omega_c})}}\cos(2\pi r\frac{E_F}{\hbar\omega_c} - \frac{\pi}{4}). \tag{25.44}$$

If we reinstate the spin splitting as well as disorder induced broadening of the DOS, the final result for the free energy becomes

$$F = F(B=0) + \frac{1}{24}(\hbar\omega_c)^2 g_0(E_F) + \frac{1}{(2\pi)^2}(\hbar\omega_c)^2 g_0(E_F)\sqrt{\frac{\hbar\omega_c}{2E_F}} \times \Big[\sum_{r=1}^{\infty}\frac{(-1)^r}{r^{5/2}} \cdot \underbrace{e^{-r\frac{\pi}{\omega_c\tau_q}}}_{P_D} \cdot \underbrace{\frac{2\pi^2 r\frac{k_b T}{\hbar\omega_c}}{\sinh{(2\pi^2 r\frac{k_b T}{\hbar\omega_c})}}}_{P_T} \cdot \underbrace{\cos(2\pi r\frac{E_F}{\hbar\omega_c} - \frac{\pi}{4})}_{P_L} \cdot \underbrace{\cos(\pi r\gamma)}_{P_s}\Big]. \tag{25.45}$$

In the above form of the free energy, the second term is proportional to $(\hbar\omega_c)^2 \propto B^2$, which gives a susceptibility $\chi_m(\text{dia.}) = -\partial^2 F/\partial B^2 = -\frac{1}{3}(\mu_B^\star)^2$, which is the non-oscillatory Landau "steady" diamagnetism, independent of B. Here $\mu_B^\star = q\hbar/2m^\star$ is the effective Bohr magneton, since it is due to the orbital motion of electrons of effective mass $m^\star$ and *not* due to electron spin. The third term is oscillatory in B: it leads to the de Haas van Alphen oscillations in the magnetic susceptibility, and also to the Shubnikov de Haas oscillations in magnetotransport as illustrated in Fig. 25.30. The sum over r converges rapidly, and often the first term $r = 1$ can give the trend reasonably accurately. We discuss the four terms indicated explicitly with underbraces:

$\mathbf{P_L}$: Is the **Landau level DOS oscillation term**, which manifests as oscillations in $E_F/(\hbar\omega_c)$, or equivalently in E_F/B in the free energy. A plot of an oscillatory physical quantity (e.g. magnetization, magnetoresistance) vs. $1/B$ will thus directly give the value of $E_F m^\star/(\hbar q)$

as the periodicity. This is used to either measure the carrier concentration, or the shape of the Fermi surface via measurement of E_F. Fig. 25.31 (a) shows how E_F is extracted from the periodicity of SdH oscillations.

$\mathbf{P_D}$: Is the **Dingle term**, which captures the damping of the oscillations due to disorder. Since this factor goes as $\exp\left[-\pi m^\star/(qB\tau_q)\right]$, the quantum scattering time τ_q can be directly measured from a plot of $\ln[P_D]$ vs. $1/B$ if the $m^\star$ is known, which is found from P_T. The Dingle plot in Fig. 25.31 (b) shows how τ_q is extracted from the decay of SdH oscillations with $1/B$.

$\mathbf{P_T}$: Is the **thermal damping term**, which decays rapidly with increasing temperature since for $\chi >> 1$, the term varies as $\chi/\sinh\chi \sim \chi e^{-\chi} \sim (2\pi^2 m^\star k_b T/\hbar qB)\cdot e^{-2\pi^2 m^\star k_b T/\hbar qB}$. Note carefully that this is *not* an Arrhenius dependence on temperature since T is in the numerator in the exponent. Because this term depends on the effective mass $m^\star$, temperature T, and magnetic field B, and not on disorder, it is used to measure the effective mass. This is done by plotting experimental values of the magnetic susceptibility or the magnetoresistance as a function of $\chi \propto T$ for various temperatures at constant B fields. It takes the shape $\chi/\sinh\chi$, from where the $m^\star$ is found by fitting the curve. Small values of $m^\star$ sustain oscillations at high temperatures, whereas for high $m^\star$ oscillation amplitudes damp rapidly with increasing temperature. Fig. 25.31 (c) shows how $m^\star$ is extracted from the decay of SdH oscillations with T. Combining P_T term with the Dingle term P_D, the temperature variation can be written as $P_D P_T \sim \exp[-C(T+T_D)/B]$, where $k_b T_D = \hbar/(2\pi\tau_q)$ defines a **Dingle temperature** T_D. For $T > T_D$ the oscillation amplitude increases sharply as the temperature is lowered. But for $T < T_D$, the amplitude does not increase further because disorder acts as an effective temperature. Of significant interest is the dependence of $P_D P_T$ on magnetic field B as $\exp[-C(T+T_D)/B]$. The deceptively interesting function $\exp[-1/x]$ **does not have a Taylor series**. This is the reason the results of quantum oscillations in magnetoresistance cannot be obtained by perturbation theory; the complete Schrödinger equation must be solved exactly, as we have done!

$\mathbf{P_s}$: Is the **spin splitting term**. It modulates the entire oscillation but does not explicitly depend on the magnetic field. Since $\gamma = gm^\star/2m_e$, it can be used to extract the g-factor if $m^\star$ is known. This term typically manifests as a spin splitting that is visible in magnetoresistance at the highest magnetic fields.

With the above discussion, we are in a position to explain experimental characteristics observed in Shubnikov de Haas oscillations. The actual calculation of the magnetoresistance fails with the Boltzmann transport equation (BTE) because the BTE uses the relaxation

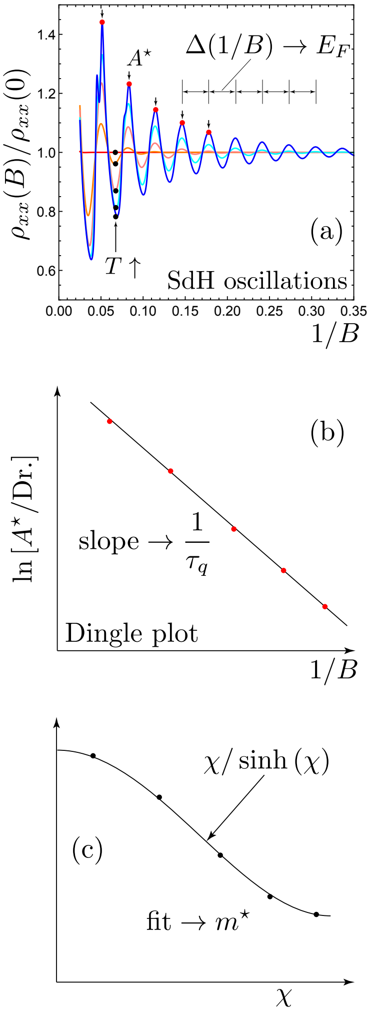


Fig. 25.31 Analysis of Shubnikov de Haas oscillations. (a) The scaled longitudinal ρ_{xx} is plotted vs. $1/B$, showing oscillations periodic in $1/B$. The periodicity $\Delta(1/B)$ yields E_F, and hence the mobile carrier concentration. (b) Dingle plot to determine the quantum scattering time. The peaks of the oscillations indicated in part (a) at a fixed temperature are the set of points $A^\star$. A plot of $\ln[A^\star/\mathrm{Dr.}]$ where Dr. $= \sqrt{\frac{\hbar\omega_c}{2E_F}}\cdot\frac{\chi}{\sinh(\chi)}$ gives the quantum scattering time τ_q. (c) A fit of the decay of the oscillation peak at a fixed magnetic field with increasing temperature as shown in (a) with an upward arrow is plotted against $\chi = 2\pi^2\frac{k_b T}{\hbar\omega_c}$ by changing $m^\star$. The best fit of the data points to the $\chi/\sinh(\chi)$ curve gives $m^\star$.

rate $1/\tau(\mathbf{k} \to \mathbf{k}')$ from Fermi's golden rule. But as our discussion above has shown, the energy eigenvalues and eigenfunctions are modified strongly by the magnetic field, and therefore the scattering rates are significantly modified. Equally importantly, the energy bandstructure is modified beyond recognition from the parabolic approximation. Therefore, one must adopt a new microscopic approach to capture the new physics of quantized magnetotransport.

This new theory was developed in the late 1950s by solving the transport problem in the density matrix formalism. We avoid getting into details of the discussion, other than pointing out that the starting point is the time-dependent Schrödinger equation itself and phase factors between initial and final states are retained for scattering events. The end result for the longitudinal magnetoresistance $\rho_{xx}(B)$ ends up being proportional to the free energy expression of equation:

$$\text{for 3D:} \boxed{\rho_{xx}(B) = \rho_{xx}(B=0)\Big[1 + \sum_{r=1}^{\infty} b_r \cos\big(2\pi r \frac{E_F}{\hbar\omega_c} - \frac{\pi}{4}\big)\Big]},$$
$$\text{where } b_r = \frac{(-1)^r}{\sqrt{r}} \sqrt{\frac{\hbar\omega_c}{2E_F}} \cdot e^{-r\frac{\pi}{\omega_c \tau_q}} \cdot \frac{2\pi^2 \frac{k_b T}{\hbar\omega_c}}{\sinh\,[2\pi^2 \frac{k_b T}{\hbar\omega_c}]} \cdot \cos\,(\pi r \gamma). \tag{25.46}$$

All the terms of Equation 25.45 carry over to the quantum magnetoresistance. Therefore, it is finally seen that all the characteristics of Shubnikov de Haas oscillations that were shown in Fig. 25.25 and listed earlier are explained by the quantum theory of Landau levels. In addition, the model also makes it clear why metals have magnetization and exhibit de Haas van Alphen oscillations in magnetoresistance, as opposed the classical Bohr–van Leeuwen theorem that did not allow any such behavior.

For 2DEGs, the same expression as Equation 25.46 is applicable, except for the absence of the $\sqrt{\hbar\omega_c/2E_F}$ factor in b_r. Writing $\chi = 2\pi^2 \frac{k_b T}{\hbar\omega_c}$, the DC and zero-$B$ conductance as $\sigma_0 = n_{2d} q^2 \tau_m / m^\star$ with τ_m the momentum scattering time, the longitudinal and transverse magnetoconductance in 2D are given by[31]

$$\sigma_{xx}(B) = \frac{\sigma_0}{1+(\omega_c\tau_m)^2}\Big(1 + 2\frac{1-(\omega_c\tau_m)^2}{1+(\omega_c\tau_m)^2} e^{-\frac{\pi}{\omega_c\tau_q}} \frac{\chi}{\sinh\chi} \cos[2\pi \frac{n_{2d}h}{2qB}]\Big),$$
$$\sigma_{xy}(B) = \frac{\sigma_0}{1+(\omega_c\tau_m)^2}\Big(1 + \frac{4}{1+(\omega_c\tau_m)^2} e^{-\frac{\pi}{\omega_c\tau_q}} \frac{\chi}{\sinh\chi} \cos[2\pi \frac{n_{2d}h}{2qB}]\Big),$$
$$\implies \rho_{xx}(B) = \rho_0\Big(1 - 2e^{-\frac{\pi}{\omega_c\tau_q}} \frac{\chi}{\sinh\chi} \cos[2\pi \frac{n_{2d}h}{2qB}]\Big), \tag{25.47}$$

where $\rho_0 = 1/\sigma_0$, the longitudinal resistivity in 2D obtained by matrix inversion. These formulae indicate that Shubnikov de Haas oscillations for the 2DEG are periodic in E_F and in $1/B$. The same method as for 3D is then used to extract the band effective mass and scattering times in 2D.

---

[31] The spin splitting terms are suppressed for clarity. A $\cos(\pi\gamma)$ term must be multiplied with the $\cos[2\pi \frac{n_{2d}h}{2qB}]$ term to account for spin.

The measurement of the quantum scattering time τ_q from the Dingle plot (Fig. 25.31) and the momentum scattering time τ_m from a measurement of the Hall effect mobility in the standard low-B Hall effect thus provides an experimental clue to the nature of the microscopic scattering times. The nature of the scattering times have a strong influence on the ratio τ_m/τ_q. As we discussed in Chapters 21 and 23, short-range scatterers such as point defects and alloy disorder lead to a ratio $\tau_m/\tau_q \approx 1$, whereas long-range scatterers such as those with Coulomb potentials lead to $\tau_m/\tau_q >> 1$.

A direct evaluation by matrix inversion from Equation 25.47 shows that the Hall resistivity ρ_{xy} in 2D does not exhibit oscillations in $1/B$. This is *not* consistent with experiments. The failure of the model is brought to sharp focus in the quantum Hall effect, which is the subject of the next section.

25.10 The quantum Hall effect

Our understanding of quantum oscillations in magnetotransport in semiconductors was thought to be mostly resolved till the remarkable discoveries of the **quantum Hall effect** in two acts. The **integer quantum Hall effect** (IQHE) was first discovered in 1980 in the 2DEG in a silicon MOSFET. The **fractional quantum Hall effect** (FQHE) was then discovered in 1982 in high mobility 2DEGs in AlGaAs/GaAs heterostructures. The models discussed in the last section are unable to explain – *even qualitatively* – these two effects. In the rest of the chapter we focus exclusively on 2DEGs in which these effects are observed.

Fig. 25.32 (a) shows a Hall-bar geometry with definitions of R_{xy} and R_{xx}. Fig. 25.32 (b) schematically shows the expected classical Hall effect in which R_{xx} increases if regions of 2DEG are removed due to localization or disorder. But $R_{xy} = B/(n_{2d}q)$, where n_{2d} is the 2DEG density, does not change with disorder. Fig. 25.32 (c) schematically shows what is actually observed at low temperatures and large magnetic fields. For low magnetic fields R_{xy} remains nearly classical, but R_{xx} develops Shubnikov de Haas oscillations as discussed in the last section. But at large fields the oscillation amplitude becomes large enough such that R_{xx} becomes zero for intervals in B. In the same intervals where $R_{xx} = 0$, the Hall resistance R_{xy} exhibits plateaus of fantastic precision, of values **exactly** equal to $R_{xy} = h/(\nu q^2) \approx 25{,}812/\nu$ Ω, **where ν is an integer, or very specific fractions**. The first observation of IQHE in Si MOSFETs showed ν to be integers. Soon a $\nu = 1/3$ step was observed in 2DEGs in AlGaAs/GaAs heterostructures. A large number of well-defined fractional values of ν have since been uncovered in samples of increasing cleanliness. The integer and fractional quantum Hall effects have been observed in a wide range of semiconductors to date, in various geometries and configurations.

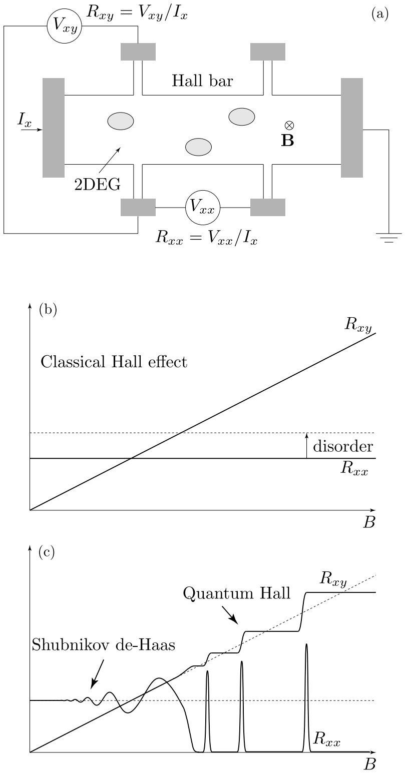


Fig. 25.32 (a) The Hall-bar geometry and definitions of currents, voltages, and resistances. The conducting channels are 2DEGs. Sometimes the entire Hall bar is gated so that the 2DEG density is tuned with gate voltage. The islands show regions from where the 2DEG electrons are removed. (b) The expected classical Hall effects, and (c) The quantum Hall effect at low temperatures and large magnetic fields.

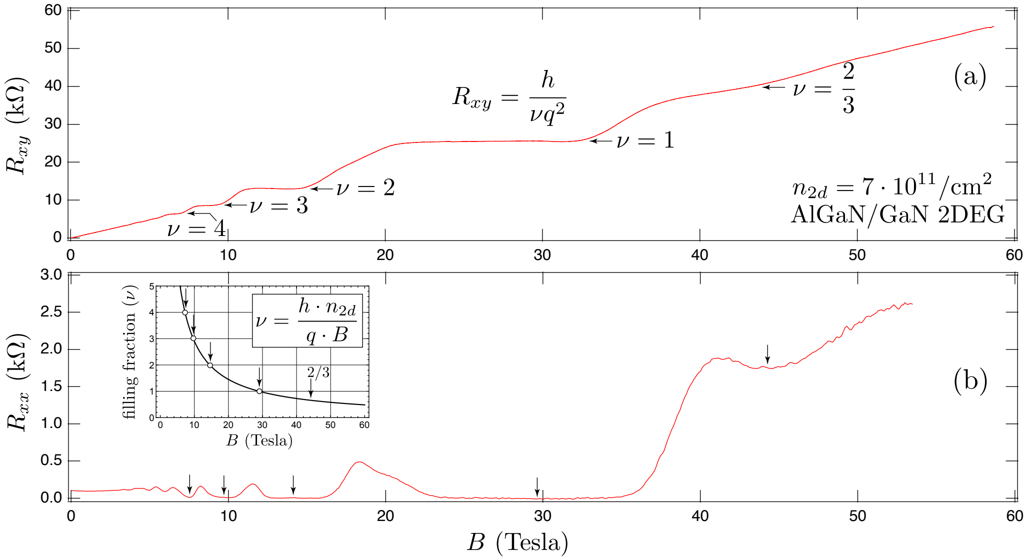


Fig. 25.33 The quantum Hall effect in an AlGaN/GaN 2D electron gas. (a) R_{xy} vs. B and (b) R_{xx} vs. B and the inset showing the filling fraction $\nu = n_{2d}/n_\phi$. Assorted data kindly shared by Scott Crooker, Los Alamos National Laboratory. See Applied Physics Letters Vol. 117, pg 262105 (2020) for details.

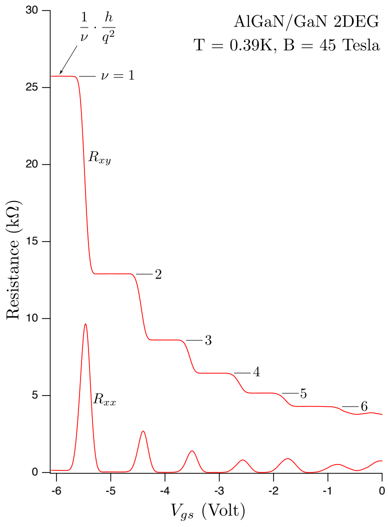


Fig. 25.34 IQHE in an AlGaN/GaN 2DEG, measured at the National High Magnetic Field Laboratory. Assorted data kindly shared by Phillip Dang and Alexey Suslov. See Science Advances, vol. 7, eabf1388 (2021) for details.

For example, Fig. 25.33 shows the quantum Hall effects in polarization induced AlGaN/GaN 2DEGs as a function of the magnetic field. The ν = integer steps marked by horizontal arrows are well resolved, and the fractional $\nu = 2/3$ step is weakly resolved. Far better resolved fractional steps are discussed later. The zeros of R_{xx} occur at **filling fractions** indicated by downward arrows. Fig. 25.34 shows the quantum Hall effect at a constant magnetic field in an AlGaN/GaN 2DEG where the 2DEG density is swept with a gate voltage. The steps of R_{xy} and the zeros in R_{xx} are well resolved. Note that the shape of R_{xx} is nearly the negative derivative of the shape of R_{xy}.

The quantum Hall effect is today "the pride" of quantum transport, and of condensed matter physics. The rather amazing fact about the quantum Hall plateaus is their **precision**, which is second only to quantum electrodynamics (QED) and fine-structure constant, which are domains of particle physics. In particle physics and QED, theory and calculations are needed to obtain the fine-structure constant. But in IQHE the fine-structure constant is **directly measured as a ratio of a voltage and a current**! The discoverer of the IQHE, K. von Klitzing (Fig. 25.5) immediately recognized the astonishing accuracy and consulted the PTB, the German standards institute for its potential for a resistance standard. The accuracy measured today in the quantized Hall resistance is nearly a part per billion. There are simply no optical, magnetic, thermal, or other measurements that come close to the accuracy of the IQHE. The fact that such accuracy is possible independent of semiconductor materials, sample geometries, and several other experimental parameters is very remarkable, and points towards deeper physics that is still in the process of being understood, as we discuss in the next section.

Even more remarkable is the discovery (Fig. 25.35) that at the face of it, it appears as if current is carried by precise *fractions* of the fundamental single electron charge[32]! The first observation of a R_{xy} plateau at $\nu = 1/3$ opened a Pandora's box of new concepts. Today the fractional quantum Hall conductance is understood to be due to the flow of not electrons, but of quasiparticles that are formed by fusing an electron with an integer number of magnetic flux quanta. These **composite fermions** exhibit several bizarre properties. Most experimentally observed fractions take the form $\nu = p/r$, where r is an odd integer. For these fractions the composite fermion theory of Jainendra Jain explains plateaus at $R_{xy} = h/(q^2\nu)$ where $\nu = \nu^\star/(2p\nu^\star \pm 1)$ where $\nu^\star$ and p are integers. Even integer fractions of $\nu = 5/2$ have also been observed, but not explained as well. The reader is encouraged to read the references at the end of the chapter to explore the rich experimental landscape of this topic. In the next section we focus on a few theories that have been devised to understand the quantum Hall effects.

Tsui
Stormer
Gossard

Fig. 25.35 D. Tsui, H. Stormer, and A. Gossard discovered the fractional quantum Hall effect. The 1998 physics Nobel prize was awarded to Tsui and Stormer (and R. Laughlin who devised the first theory to explain this phenomena).

25.11 Quantum Hall effect theories

A theoretical model of the quantum Hall effect must transcend the picture of scattering limited transport to explain the precision of quantization, and simultaneously account for the role of scattering[33]. There are theories that explain why plateaus and zeroes of high precision should occur. But there is no complete theory to date that provides closed expressions for the quantized R_{xy} and R_{xx} that explain the observed integer and fractional quantum numbers, the temperature dependence, and the plateau widths (e.g. similar to those of Shubnikov de Haas). Since the search for such a theory is still a subject of research – in this section we take an experimental evidence driven approach to discuss our current understanding. To do so, we first collect several definitions discussed earlier in the chapter, but adapt them for 2D electron systems to streamline the discussion that will follow.

Landau level energies are $E_n = (n + \frac{1}{2})\hbar\omega_c$ where $n = 0, 1, 2, ...$ for a 2DEG. Note the absence of kinetic energy since the (k_x, k_y) motion is "quenched" in the bulk because of cyclotron motion in the $x - y$ plane in the interior of the 2DEG. But as we shall see soon, propagating edge states can exist.

Broadened DOS: As shown in Fig. 25.23 (b), the DOS of a 2DEG in the absence of disorder is a series of Dirac-δ functions at the Landau level energies $E_n = (n + \frac{1}{2})\hbar\omega_c$ with a degeneracy $\hbar\omega_c \cdot \frac{m^\star}{\pi\hbar^2}$ for a spin degeneracy of $g_s = 2$, neglecting Zeeman splitting. In the presence of disorder, the DOS broadens. The broadening is critical to the explanation of the quantum Hall effect.

---

[32] To quote from Stormer's Nobel Lecture in the Reviews of Modern Physics, Vol. 71, No. 4, pg. 875 (1999): "*...Lightheartedly, Dan Tsui enclosed the distance between $B = 0$ and the position of the last IQHE (~7 cm) between two fingers of one hand and measured the position of the new feature in this unit. He determined it to be three and exclaimed, "quarks!" Although obviously joking, with finely honed intuition, he had hit on the very essence of the data.*"

[33] A note of caution that this rather long section is written in the style of a summary of a fascinating topic that still has unanswered questions. It should be read in that spirit, and appended with the references at the end of the Chapter.

Magnetic length: $l_B = \sqrt{\hbar/qB}$ is the spatial extent of the electrons in the Landau eigenstates in the $x - y$ plane. The center of the orbit of the eigenstate in the Landau gauge $\mathbf{A} = B(0, x)$ and $\mathbf{B} = B\hat{z}$ is $x_{k_y} = -k_y l_B^2$, and $k_y = (2\pi/L_y)n_y$ where $n_y = 0, \pm 1, ...$

Flux quantum: The magnetic flux quantum is $\phi_0 = h/q$. Note that $\phi_0 = B \cdot L_y \cdot (x_{k_y} - x_{k_y - \frac{2\pi}{L_y}}) = B \cdot L_y \cdot \frac{\hbar}{qB} \cdot \frac{2\pi}{L_y} = \frac{h}{q}$. This is illustrated in Fig. 25.22. It simply means that for a 2DEG there is one flux quantum per allowed Landau level eigenstate for electrons.

Flux density n_ϕ: At a magnetic field B, the density of flux lines piercing the 2DEG plane is $n_\phi = B/\phi_0$, which has units of cm^{-2}. We have seen earlier that there is one flux quantum for every allowed Landau eigenstate for electrons, obtained as the quantized solution of the Schrödinger equation for 2DEG electrons in a magnetic field. Therefore n_ϕ is also the degeneracy of a Landau level in 2D.

Filling factor ν: The number of electrons per flux quantum, $\nu = \frac{n_{2d}}{n_\phi} = \frac{h \cdot n_{2d}}{q \cdot B}$, where n_{2d} is the density of electrons in the 2DEG.

The Aharonov–Bohm phase: For a circular orbit in a magnetic field, a Bohr–Sommerfeld type quantization condition requires that both the circular orbit due to boundary conditions, as well as the cyclotron orbit due to the magnetic field must close on themselves[34]. Writing the de-Broglie relation as $m^\star \mathbf{v} = \hbar \mathbf{k}$, for a path around an orbit of radius R_c, this requires $e^{i \oint (\mathbf{k} - \frac{q}{\hbar}\mathbf{A}) \cdot d\mathbf{l}} = e^{i2\pi n}$ where n is an integer. The **dynamical phase factor** is obtained from the semiclassical relation $\hbar k = m^\star v$ with $v = \omega_c R_c$ to give $e^{i \oint \mathbf{k} \cdot d\mathbf{l}} = e^{ik2\pi R_c} = e^{i \cdot \frac{m^\star \omega_c R_c}{\hbar} \cdot 2\pi R_c} = e^{i4\pi \frac{\phi}{\phi_0}}$, where $\phi = B \cdot \pi R_c^2$ is the magnetic flux through the orbit. The part due to $\mathbf{A}$ is the **Aharonov–Bohm phase factor**: $e^{-i \oint \frac{q}{\hbar} \mathbf{A} \cdot d\mathbf{l}} = e^{-i2\pi \frac{\phi}{\phi_0}}$, where by Stoke's theorem $\oint \mathbf{A} \cdot d\mathbf{l} = \int \mathbf{B} \cdot d\mathbf{S} = \phi$ is the magnetic flux through the orbit[35]. The sum of the dynamical and Aharonov–Bohm phases must therefore be $4\pi \frac{\phi}{\phi_0} - 2\pi \frac{\phi}{\phi_0} = 2\pi \frac{\phi}{\phi_0} = 2\pi n$. This means the **total magnetic flux through a closed electron orbit must be an integer multiple of the magnetic flux quantum** ϕ_0. Note the Aharonov–Bohm phase is half of the dynamical phase in magnitude.

Let's now explore the mechanisms that are currently believed to cause the quantum Hall effects. Several treatments of the subject begin the theoretical discussion with an emphasis on the integer quantum Hall effect. This makes the fractional quantum Hall effect appear mysterious and fundamentally different. Instead, we will treat them on the same footing using the **composite fermion** concept[36]. The best justification to do so is to observe carefully some remarkable symmetries hiding in plain sight in the experimental data shown in Fig. 25.36 (b) for high mobility AlGaAs/GaAs 2DEGs. First, note that the fractional plateaus and zeroes (e.g. $\nu = 1/3$) are as well developed as the integral ones (e.g. $\nu = 1$). A line below the x-axis marked ⓐ indicates in gray boxes the zeroes (and plateaus) starting from $B = 0$. Note carefully that if this line is translated to the $\nu = 1/2$ point along the x-axis to

---

[34] In the presence of a magnetic field, the total momentum for an electron is $\mathbf{p} = m^\star \mathbf{v} - q\mathbf{A}$ where $m^\star \mathbf{v}$ is the kinetic and $-q\mathbf{A}$ is the "potential" momentum. Thus the kinetic energy which enters the Schrödinger equation is $(\mathbf{p} + q\mathbf{A})^2/2m^\star$.

[35] The Aharonov–Bohm phase has the remarkable consequence that the phase of the electron is affected by a magnetic field B even when $B = 0$ where the wavefunction of the electron is non-zero! The classic experiment is where electrons wind around two paths of a split ring pierced by a solenoid. Though $B = 0$ in the path of electrons, they still exhibit interference patterns in their transport properties as a function of B because $\mathbf{A} \neq 0$ in their path.

[36] For several years the fractional steps remained mysterious till the idea of composite fermions was introduced by J. Jain in Physical Review Letters, vol. 63, page 199 (1989).

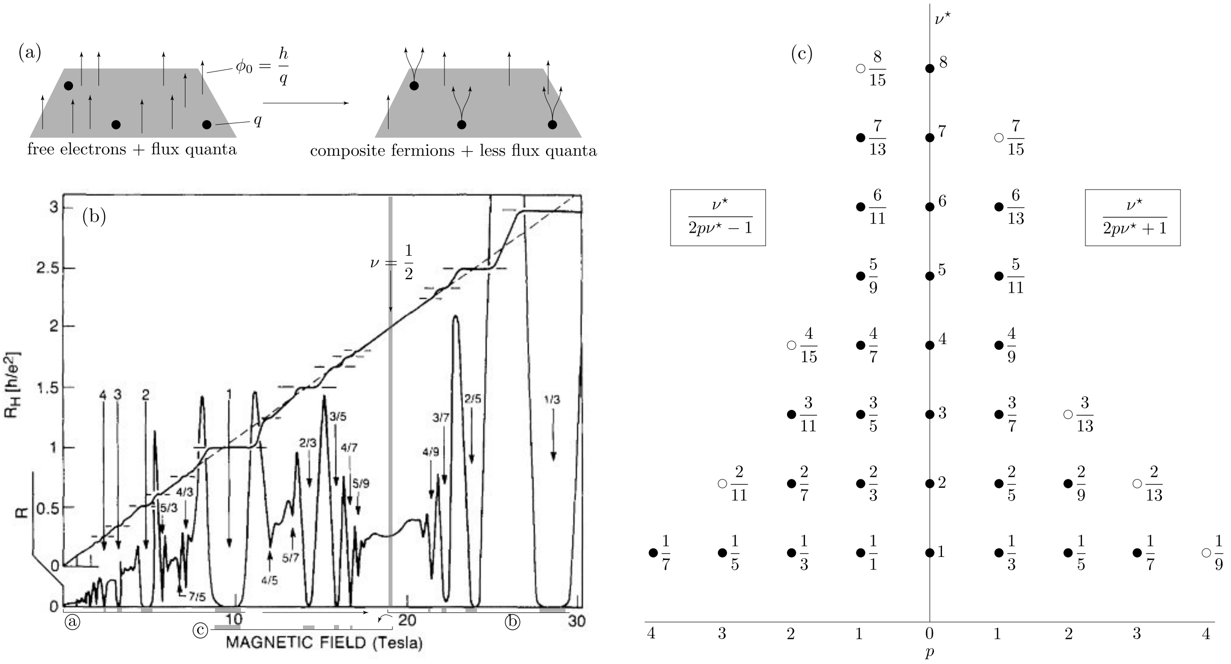


Fig. 25.36 (a) An illustration of composite fermions. (b) Experimental quantum Hall effect from Physica B vol. 177 page 401 (1992). (c) Fractions and schematic representation of composite fermions adapted from Physical Review Letters, vol. 63, page 199 (1989).

ⓑ, the zeroes and plateau locations match up again! Even more interestingly, if the line ⓑ is reflected about the $\nu = 1/2$ line to the line ⓒ, the zeroes and plateaus again coincide. It seems as if the effective magnetic field $B^\star$ becomes zero for $\nu = 1/2$, the filling fraction when for each electron there are exactly 2 flux quanta ϕ_0.

This repeating pattern is characteristic of the quantum Hall effect. It indicates that the transport is not of bare electrons, but of quasiparticles in which an even number of flux quanta are attached to each electron, as indicated schematically in Fig. 25.36 (a). If each electron "swallows" an even number of flux quanta $2p\phi_0$ where p is an integer, the quasiparticle[37] experiences an effective magnetic field $B^\star = B - 2pn_{2d}\phi_0$. This yields a quasiparticle filling fraction $\nu^\star$ via

$$B^\star = B - n_{2d}\cdot(2p\phi_0) \implies \frac{n_{2d}\phi_0}{\nu^\star} = \frac{n_{2d}\phi_0}{\nu} - 2pn_{2d}\phi_0$$

$$\implies \frac{1}{\nu^\star} = \frac{1}{\nu} - 2p \implies \boxed{\nu = \frac{\nu^\star}{2p\nu^\star \pm 1} \text{ and } R_{xy} = \frac{h}{\nu q^2}.} \qquad (25.48)$$

The $-$ve sign in the box is if $B^\star$ points in the direction opposite to B. The fractions obtained for integer values of $\nu^\star$ and p are shown in Fig. 25.36 (c). Note that these are indeed the R_{xy} plateaus and corresponding R_{xx} zeroes observed in the experiment in Fig. 25.36 (b). The above model seems plausible if the composite particle of [1 electron + $2p\phi_0$ flux quanta] is indeed a fermion.

---

[37] A crude analogy is an oxygen atom capturing two hydrogen atoms to form water molecules to lower the total energy. The motion of the water molecule is different from the constituent atoms. Yet another comparison is to a Cooper pair in superconductors, in which two electrons pair via a phonon. Here, two flux quanta pair via an electron.

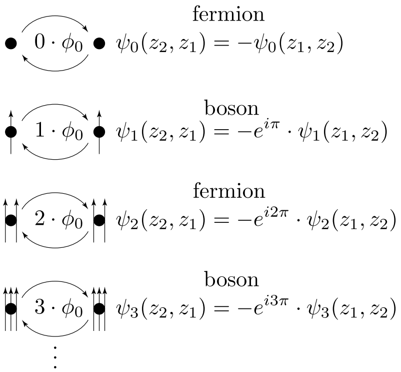

fermion
$0 \cdot \phi_0$
$\psi_0(z_2, z_1) = -\psi_0(z_1, z_2)$
boson
$1 \cdot \phi_0$
$\psi_1(z_2, z_1) = -e^{i\pi} \cdot \psi_1(z_1, z_2)$
fermion
$2 \cdot \phi_0$
$\psi_2(z_2, z_1) = -e^{i2\pi} \cdot \psi_2(z_1, z_2)$
boson
$3 \cdot \phi_0$
$\psi_3(z_2, z_1) = -e^{i3\pi} \cdot \psi_3(z_1, z_2)$
⋮

Fig. 25.37 Quantum statistics of composite fermions.

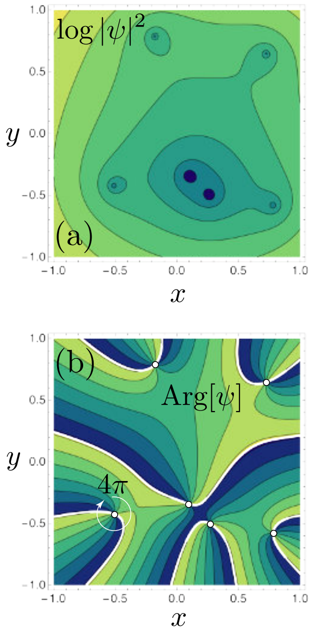


Fig. 25.38 Schematic representation showing (a) the amplitude of the many-particle wavefunction of an electron in the quantum Hall state $\nu = 1/2$. The amplitude is zero at the location of another electron. (b) The phase of the many-particle wavefunction for $\nu = 1/2$. If an electron is taken around any other electron in the 2D plane, its phase changes by $2\pi/\nu = 4\pi$. This is equivalent to the case if 2 flux quanta were attached to each electron at $B = 0$.

Consider two electrons in the 2D plane as shown in Fig. 25.37. Let the location of an electron in the (x, y) plane be denoted by the single dimensionless complex number $z = (x - iy)/l_B$. Then, the effect of swapping the location of two electrons when $B = 0$ is that the new wavefunction $\psi_0(z_2, z_1)$ is the same as the old wavefunction $\psi_0(z_1, z_2)$ except for a negative sign: this is the fingerprint of electrons being fermions as discussed in Chapter 3.

But now consider an electron that has "captured" one magnetic flux quantum ϕ_0 indicated by an arrow. If we swap this composite object, each electron experiences half the flux quanta, and the combined wavefunction acquires another negative sign due to the Aharonov–Bohm phase of $\exp\left[-i2\pi\frac{\phi_0/2}{\phi_0}\right] = \exp[-i\pi] = -1$. This composite object then follows bosonic statistics since the wavefunction is unaltered by swapping $z_1 \leftrightarrow z_2$. If the electron captures two flux quanta $2\phi_0$, the swap flips the sign, thus it follows fermionic statistics again. Extending this argument indicates that if an electron captures an **even number** of flux quanta, the combined object of [1 electron + even flux quanta] indeed behaves as a **composite fermion**.

The above plausibility argument is strengthened if the total energy of the composite fermion is found to be lower than if the electron did not capture the flux quanta. This has been confirmed by writing down the many-particle wavefunction of the composite fermion, feeding it into the Schrödinger equation, and showing that the total energy is indeed lowered because of the formation of the composite fermion. This test is performed numerically. The reason for the lowering of energy is the reduction of Coulomb repulsion since most of the kinetic energy is quenched in the high magnetic field.

The composite fermion wavefunction at a filling fraction ν and effective fraction $\nu^\star$ is of the form $\psi_\nu = \Pi_{j<k}(z_j - z_k)^2 \mathcal{S}_{\nu^\star}$, where z_j, z_k are the (complex) locations of electrons j and k, and $\mathcal{S}_{\nu^\star}$ is a Slater determinant formed of the single-particle wavefunctions of the Landau level states. For example, Figures 25.38 (a) and (b) simply show how the amplitude and phase of $\Pi_{j<k}(z_j - z_k)^2$ without $\mathcal{S}_{\nu^\star}$ varies in space for several electrons located randomly in the 2D plane. Note in this figure that a closed path in Fig. 25.38 (b) around any electron in the 2D plane seen as a node in Fig. 25.38 (a) causes a total phase change of $2\pi/\nu$, which is 4π for $\nu = 1/2$. This is the same phase as due to two flux quanta. Let us investigate the more general case.

Now consider taking an electron around a path of area A in the 2D plane in the general case. When one electron is taken in a closed loop around a single other electron assumed at the origin, the factor z^2 in the wavefunction produces a phase change of 4π. But if the 2DEG density is n_{2d}, all electrons in the enclosed area A cause a total dynamical phase change of $4\pi n_{2d} A$. The corresponding Aharonov–Bohm phase due to a magnetic field B is $-2\pi BA/\phi_0$. This makes the total phase change $\Delta\phi = 4\pi n_{2d} A - 2\pi BA/\phi_0 = 4\pi n_{2d} A - 2\pi n_\phi A$. If the entire phase is assumed to be the Aharonov–Bohm phase due to an **effective magnetic field** $B^\star$, then we get $-2\pi\frac{B^\star A}{\phi_0} = 4\pi n_{2d} A - 2\pi\frac{BA}{\phi_0} A$, which confirms what we gleaned from the experiment: $B^\star = B - 2n_{2d}\phi_0$.

Additional experimental evidence of composite fermions is obtained from a) measurement of its effective mass, b) the temperature dependence of the resistivity ρ_{xx} for fractional ν, c) cyclotron resonance, and d) evidence for their collective orbital motion. The reader is directed to the references at the end of the chapter for further exploration.

The composite fermion picture, therefore, to an extent, removes the distinction between the integer and fraction quantum Hall features and gives a map of the magnetic field values at which the Hall resistance R_{xy} and the longitudinal resistance R_{xx} exhibit quantized features. We are now ready to discuss the heart of the matter: why are there **plateaus** of R_{xy} and **zeroes** of R_{xx} at the steps $\nu = \nu^\star/(2p\nu^\star \pm 1)$? As we will see now, the quantum Hall effect offers a "catch-22" sort of situation in which **disorder** is necessary for the fantastic precision of the plateaus in R_{xy}, but too much disorder washes out the plateaus completely!

Let us first discuss a few unconventional features that do not require new concepts, but are simply the consequences of a) the formation of Landau levels and the modified electron density of states in a large magnetic field, and b) the tensor relation between current and electric fields. First note that in the experimental data shown in Fig. 25.33 (a), between the quantum Hall effect plateaus, the Hall resistance R_{xy} follows a classical linear dependence on the magnetic field, the classical Hall effect $R_{xy} = B/(n_{2d}q)$. In Fig. 25.33 (a), as well as in Fig. 25.34, the Hall resistance R_{xy} goes through plateaus at $R_{xy} = \frac{1}{\nu}\frac{h}{q^2}$ where $\nu = hn_{2d}/qB$ is the filling fraction. The behavior of the longitudinal resistance R_{xx} is intriguing: at the values of magnetic field where R_{xy} forms plateaus, $R_{xx} \to 0$.

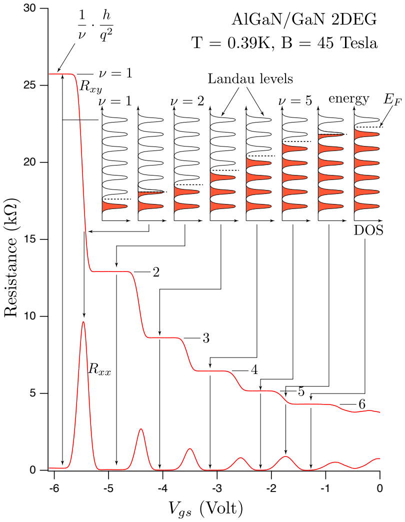


Fig. 25.39 Explanation of Fig. 25.34 showing the correspondence of R_{xy} and R_{xx} with the Landau level DOS occupation and the Fermi level. Note that when the Fermi level is in the gaps between Landau levels, R_{xx} goes to zero and R_{xy} exhibits plateaus. But when E_F is inside and near the middle of a Landau level, R_{xx} and R_{xy} do not exhibit zeros and steps respectively, they behave more like a classical metallic system exhibiting features similar to the classical Hall effect.

Fig. 25.39 is a first attempt at explaining this behavior. When the Fermi level E_F is in the middle of a Landau level, there is a finite longitudinal resistance R_{xx} and the Hall resistance R_{xy} is not quantized. The quantization of R_{xy} and the vanishing of R_{xx} occurs when the Fermi level is in the gaps in the DOS between Landau levels. Note that while R_{xy} measures *all* carriers – in every Landau level, R_{xx} only depends on the carriers at the Fermi level E_F. Because $\nu = \frac{n_{2d}}{n_\phi} = \frac{n_{2d}h}{Bq}$, $R_{xy} = \frac{B}{n_{2d}q} = \frac{h}{\nu q^2}$, which only states that R_{xy} should take the values $\frac{h}{\nu q^2}$ at integer and fractional ν. It does not offer any explanation of why R_{xy} actually forms *plateaus*. The explanation of plateaus requires a discussion of disorder and localization, to which we will return after a discussion of the mystery of the vanishing R_{xy}.

Our intuition from the scalar nature of resistance says that R_{xx} should be *high* when the Fermi level is in a gap, and it should be *low* when E_F is in a high DOS region. But in Fig. 25.39 the behavior is upside down! The vanishing of the longitudinal resistance R_{xx} may be naively interpreted as a very high conductance state – or even a superconducting state. However, it is the exact opposite – the longitudinal conductivity state is not a superconductor, and not even a metallic state: it is an *insulating* state! To see this, we must remember that the resistivity ρ and the conductivity σ in the presence of a large magnetic field are **tensor** quantities. For a 2D system the σ matrix is defined by:

$$\begin{bmatrix} J_x \\ J_y \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \end{bmatrix}, \tag{25.49}$$

and the resistivity matrix ρ is the inverse of the conductivity matrix:

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} \rho_{xx} & \rho_{xy} \\ \rho_{yx} & \rho_{yy} \end{bmatrix}^{-1} = \frac{1}{\rho_{xx}\rho_{yy} - \rho_{xy}\rho_{yx}} \begin{bmatrix} \rho_{yy} & -\rho_{xy} \\ -\rho_{yx} & \rho_{xx} \end{bmatrix}. \tag{25.50}$$

For the QHE plateaus, $\rho_{xx} = \rho_{yy} = 0$, which implies that $\sigma_{xx} = 0$. In between the QHE plateaus, ρ_{xx} is finite and therefore therefore σ_{xx} is metallic. So the longitudinal conductivity σ_{xx} goes through a series of metal-insulator transitions, which follows from the movement of the Fermi-level through the oscillating density of states. The metallic longitudinal conductivity depends on scattering and impurities, and does not exhibit any sort of precise quantization as does the longitudinal conductivity.

A hint towards understanding the vanishing of R_{xx} and the plateaus in R_{xy} is obtained from the following classical analysis. The simultaneous effect of a voltage and a very large magnetic field on a 2D conductor is shown in Fig. 25.40. The plots are of the electrical potential profiles obtained by numerical solution inside a 2D conducting rectangular bar at (a) zero magnetic field, and (b) a very large magnetic field. A 1 volt potential difference is applied along the x-direction. In the absence of a magnetic field, Fig. 25.40 (a) shows a linear potential drop from 1 volt to 0 volt as is expected along the x-direction with no dependence on the y-direction at all. If voltage probes were connected at the three points shown as circles, then V_{xx} will depend on the distance between the probes, but $V_{xy} = 0$: it is a simple conductor.

Fig. 25.40 (b) shows the potential profile in a very large magnetic field. The net result is that the voltage along the left edge of the bar is shorted to $V(x, 0) = 0$, and the opposite edge of the bar is shorted to $V(x, L_y) = V$. Thus, a measurement of $R_{xx} = V_{xx}/I$ on either edge of the conductor would yield zero, *irrespective of where the voltage probes are placed*. R_{xy} will be proportional to the entire V! There are also sharp potential drops at opposite corners of the bar. The color scale indicates equipotential lines: they curve from the top left corner to the bottom right corner through the middle of the bar, but equipotential lines near the edges of the conductor are parallel to the edge. This behavior of equipotential lines is responsible for the flow of **edge currents** in the quantum Hall states, and to localized heating at the bottom right and top left corners of sharp potential gradients.

Fig. 25.40 (c) illustrates the **edge states** in real space. If electrons enter into the conductor and move *without losing energy*, they must move along equipotential lines. Since they are entering from opposite electrodes with energies corresponding to $V = 0$ and $V = V$, they 1)

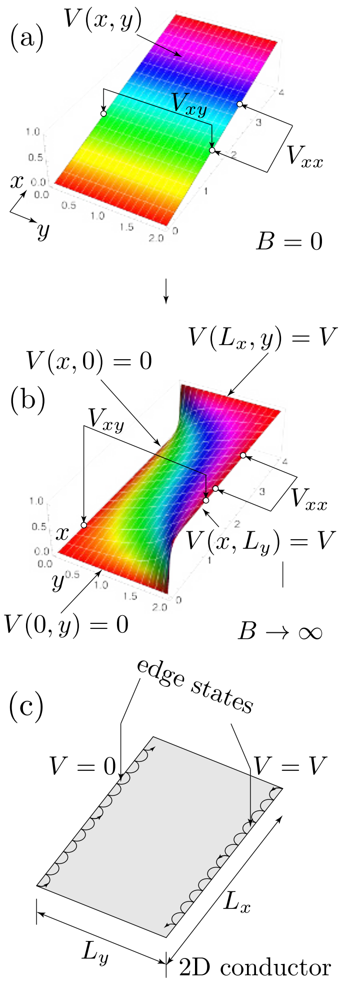


Fig. 25.40 Potential distribution in response to a voltage difference in the x-direction inside a 2D conducting rectangular bar of size $(L_x, L_y) = (2, 4)$ in arbitrary length units at (a) $B = 0$ showing a linear voltage drop V_{xx} along x and zero lateral voltage drop V_{xy} along y. (b) The potential profile at a very large magnetic field applied in the z-direction. Note that for very large B, $V_{xx} = 0$ on both left and right edges. In addition, $V_{xy} = V$ is the *entire* voltage applied in the x-direction. (c) Real space trajectories of electrons in the conductor at large B. Electrons entering the conductor without losing energy are allowed to enter from a separate electrode on each edge, remain confined to the edge after they enter, and only move in one direction: they are **chiral**.

must enter from a separate electrode into each edge, 2) must remain confined to the edges, and 3) can only move in one direction: they are **chiral**. The left edge shares the same potential as the electrode with $V = 0$, whereas the right edge shares the same potential with the electrode of potential V. The **chiral** motion of the skipping orbits means on the right edge electrons move only along the $-x$ direction, whereas on the left edge they only move in the $+x$ direction. The crucial effect of the B field is to split these chiral modes widely in space: they are separated by the width of the conductor L_y. For large enough magnetic field, the radius of the skipping orbits R_c is much smaller than the sample width L_y. When this condition $R_c << L_y$ is satisfied, the chiral edge modes become impervious to roughnesses at the edges: they simply cannot back-scatter because then they have to jump to the other edge of the conductor which is prevented by the small orbit. They will also then require to change energy, which is suppressed in the absence of inelastic scattering mechanisms.

Fig. 25.41 (a) shows a Hall bar with a 2DEG and Fig. 25.41 (b) shows the energy levels variation along the width of the bar at a high magnetic field. The energy level variation in Fig. 25.41 (b), which we crudely refer to as the **Landau band diagram**, is the quantum version of Fig. 25.40. Before we outline the method of calculation, we discuss the end result. At a high magnetic field, the energy eigenvalues of electron states with the center of cyclotron orbits far from the edges form Landau levels of energy $E_n = (n + \frac{1}{2})\hbar\omega_c$ as discussed earlier. When the Fermi level is in between these "flat-band" regions, the flat-band energy states do not contribute to σ_{xx}, which goes to zero. But near the edges, the confinement potential combined with the magnetic field produces an extra effective potential. The potential $V_{edge}(y)$ produces a net Lorentz-like force $-\partial V_{edge}(y)/\partial y = qv_x B$. Since the potential pushes electrons inwards from the edge towards the center or the bar, the net Lorentz force points along the edge in the $+x$ direction in one edge and in the $-x$ direction in the other edge. These 1D edge states carry a quantum of conductance from the Landauer–Buttiker formulation: the number of 1D edge modes is the filling factor ν (= 2 in Fig. 25.41), leading to the integer quantum Hall conductance $\sigma_{xy} = \nu q^2/h$. The 1D edge states were introduced by B. Halperin in 1982 shortly after the discovery of the quantum Hall effect.

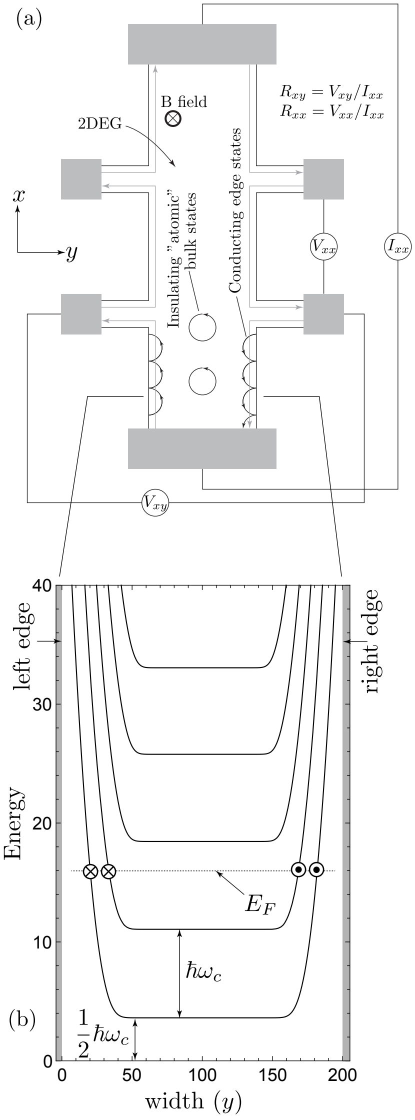


Fig. 25.41 Illustration of the Hall-bar (a) and (b) Landau level energy "bands" along the width of the bar. The confinement potential at the sample edge bends the energy levels upwards, similar to the upward surface band bending in the depletion region in a n-type doped semiconductor. In the doped semiconductor the band bending is due to a real electric field due to depleted space charges, but in the Hall bar the "edge band-bending" is due to the Lorentz-force induced potential.

We now discuss the quantum model of transport of electrons in crossed electric and magnetic fields, and then introduce disorder into the quantum model to explain the quantum Hall plateaus of R_{xy} and zeroes of R_{xy}. As a first step, let us update the classical model of the Hall effect discussed in early sections of this chapter. We now solve the quantum theory of the standard Hall effect by adding an electric field to the magnetic vector potential of Equation 25.30 for an electron in a periodic potential without disorder. This analytically solvable quantum problem yields the same Hall effect current as the classical result!

Following the axes of Fig. 25.40 and Fig. 25.41, the electric field vector is $\mathbf{E} = (-E_x, 0, 0)$, leading to an electric potential $V(x) = qE_x x$, which adds to the magnetic vector potential. As in the classical case, we define the Hall velocity as $v_H = E_x/B$, and $x_0 = (\hbar k_y + m^\star v_H)/(qB)$. Carrying over the results from Equation 25.30 and adapting for 2D by removing k_z, the Schrödinger equation and its solution for the crossed **E** and **B** fields are:

$$\Big(\frac{1}{2m^\star}\big(\begin{bmatrix}\hat{p}_x\\ \hat{p}_y\end{bmatrix} + q\begin{bmatrix}0\\ Bx\end{bmatrix}\big)^2 + qEx\Big)\psi = E\psi, \text{ where } \psi = e^{ik_y y}\phi(x) \text{ gives}$$

$$\Big(\frac{\hat{p}_x^2}{2m^\star} + \frac{m^\star\omega_c^2}{2}(x+x_0)^2 + x_0\cdot qE_x + \frac{1}{2}m^\star v_H^2\Big)\phi = E\cdot\phi$$

$$\boxed{E = (n+\frac{1}{2})\hbar\omega_c + qE_x x_0 + \frac{1}{2}m^\star v_H^2} \;\&\; \boxed{\psi_{n,k_x}(x,y) = \frac{e^{ik_y y}}{\sqrt{L_y}}H_n(x+x_0).} \tag{25.51}$$

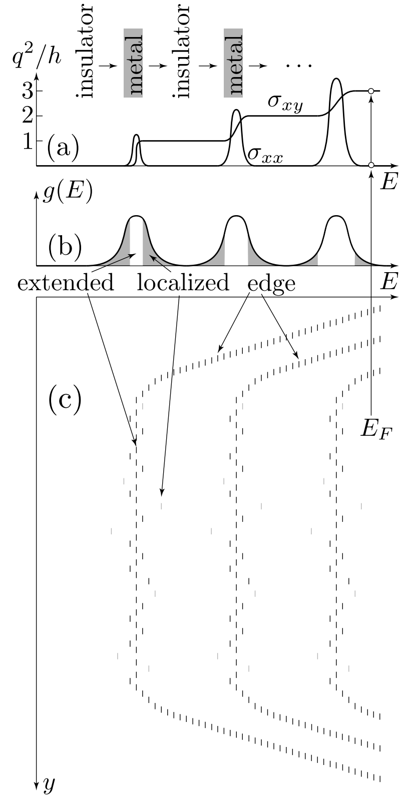


Fig. 25.42 Illustration of the effect of disorder on the Landau levels, the energy eigenvalue dispersion, and the resulting effect on the quantum Hall conductance steps.

The electric field reduces the degeneracy of the Landau levels through the extra field-dependent terms $qE_x x_0 + \frac{1}{2}m^\star v_H^2$, and now becomes linearly dependent on k_y via x_0. The group velocity is therefore $\frac{1}{\hbar}\nabla_{\mathbf{k}}E = v_H\hat{y} = \frac{E_x}{B}\hat{y}$, the classical Hall velocity in the direction $\mathbf{E}\times\mathbf{B}$. The center of the orbits are at x_0, which depend on k_y. In other words, the energy eigenvalues can be plotted either as a function of x_0 along the electric field, or alternatively as a function of k_y as a "bandstructure". The motion of carriers along the y-axis leads to the Hall voltage.

To explain the plateaus of σ_{xy} and zeroes of σ_{xx}, we now introduce both an edge confinement potential $V_{edge}(y)$ independent of x, as well as a disorder potential $V_{disorder}(x,y)$ into the above Schrödinger equation. Without disorder, the edge confinement potential leads to Fig. 25.41 (b), and then disorder modifies Fig. 25.41 (b) to the diagrams shown in Fig. 25.42 and Fig.25.43. Note that the disorder due to width variations may be absorbed into $V_{disorder}(x,y)$.

Addition of the edge potential converts the Schrödinger equation and its solution into

$$\Big(\frac{\hat{p}_x^2}{2m^\star} + \frac{m^\star\omega_c^2}{2}(x+x_0)^2 + x_0\cdot qE_x + \frac{1}{2}m^\star v_H^2 + V_{edge}(y)\Big)\phi = E\cdot\phi$$

$$E(y_0) = (n+\frac{1}{2})\hbar\omega_c + qE_x x_0 + \frac{1}{2}m^\star v_H^2 + \tilde{V}_{edge}(y_0), \tag{25.52}$$

where $\tilde{V}_{edge}(y_0)$ is evaluated as the matrix element $\langle\psi|V(y)|\psi\rangle$ for the Landau wavefunction centered at $y = y_0 = -k_y l_B^2$. Clearly for y far from the edges of the Hall bar, $V(y) = 0$ and therefore the allowed eigenvalues are simply $E = (n+\frac{1}{2})\hbar\omega_c$ at a small E. In the absence of a disorder potential, these eigenvalues far from the edges are indeed independent of y and of k_y, as illustrated in Fig. 25.41 (b). The velocity of the states $\frac{1}{\hbar}\nabla_{\mathbf{k}}E$ therefore is zero in the bulk, but at the edges for the n^{th} Landau level band for $E \approx 0$ it is

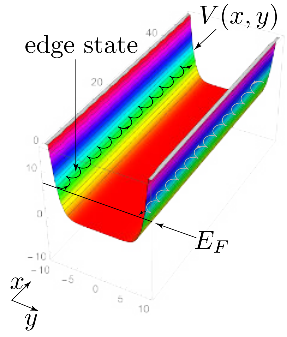
$$v_n(k) = \frac{1}{\hbar}\nabla_k E(y_0) = \frac{1}{\hbar}\frac{\partial\tilde{V}_{edge}(y)}{\partial k_y} = \frac{1}{\hbar}\frac{\partial\tilde{V}_{edge}(y)}{\partial y_0}\frac{\partial y_0}{\partial k_y} = \frac{1}{qB}\Big(-\frac{\partial\tilde{V}_{edge}(y)}{\partial y_0}\Big), \tag{25.53}$$

which is in the form of a Lorentz force. It is important to note though that the exact velocity of an edge state *cancels* when the net current is evaluated because the edge state is a 1D state, for which the density of states is inversely proportional to the group velocity. The current is

$$I = q \int \frac{dk}{2\pi} \frac{1}{\hbar} \frac{dE}{dk} f(E) = \frac{q}{2\pi\hbar} \int_{\mu_L}^{\mu_R} dE f(E) = \frac{q}{h}(\mu_R - \mu_L) = \frac{q^2}{h} V \tag{25.54}$$

for each edge mode, irrespective of the shape of $V_{edge}(y)$! From Fig. 25.42 (c), if the Fermi level E_F is at a location above 3 Landau level floors intersecting the upward curved parts, the bulk states $n = 0, 1, 2$ are filled completely and cannot carry a net current in the x direction by Pauli exclusion, giving $\sigma_{xx} = 0$. But there are $\nu = 3$ edge states, and the Hall conductivity from Equation 25.54 is $\sigma_{xy} = 3q^2/h$, as indicated by the arrows and circles vertically above the E_F.

(a) without disorder



(b) with disorder

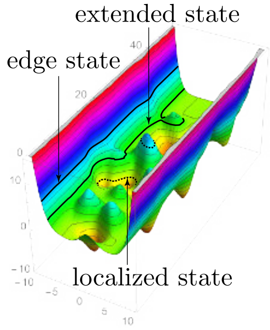


Fig. 25.43 Illustration of the net potential experienced by electrons in a large magnetic field, edges, and disorder.

Let us consider now the situation when E_F is lowered with a gate voltage. The quantum Hall plateau $\sigma_{xy} = 3q^2/h$ and the zero σ_{xx} stay constant, till E_F reaches the flat-part of the Landau levels. The addition of a disorder potential $V_{disorder}(x, y)$ will shift the local eigenvalues as indicated in Fig. 25.42 (c). For different values of n, the energy eigenvalues thus take a disordered bathtub short of shape. Fig. 25.42 (b) shows the corresponding density of states.

It is clear from this argument that there are states in the flat bottom part that can be connected across along the current direction, which are called the extended states. There are also states that are localized due to a stronger disorder potential. At the centers of each broadened Landau levels are the extended states, and the localized states are at the flanks.

Since the disorder potential also varies along the current flow, or x-direction, the situation is illustrated in Fig. 25.43 (a) for the case without disorder, and in 25.43 (b) for the case with disorder. In Fig. 25.43 only one of the Landau bands is shown. Fig. 25.43 (a) shows the edge states at the opposite edges as skipping orbits, and Fig. 25.43 (b) shows an edge state center of orbit motion as a line.

When E_F is lowered into the Landau level peak in DOS, extended and localized states co-exist. The extended states lead to a finite σ_{xx}, as indicated by a connected line in Fig. 25.43 (b) that goes around defects in the current direction. The localized states around hills and valleys are also indicated by dashed loops in the this figure. Thus, the smaller is the disorder, the sharper is the transition from one plateau of σ_{xy} to the next. Fig. 25.42 (a) thus indicates that the quantum Hall steps can be considered to be a series of transitions from insulating to metallic bulk states. Every time there is a transition, the number of edge states changes by unity.

In the discussion till now, the precise role of the electrodes has not been considered. The role of electrodes is now understood to be central to the quantum Hall effect. An electrode that is a current source or sink injects electrons into a conductor at all energies up to its Fermi level. A voltage probe does not sink or source electrons, or energy. A voltage probe is an elastic scatterer: if an electron enters it, another comes back with the same energy with a randomized momentum. Therefore, a voltage probe preserves the energy of a beam of electrons bouncing off from it.

Consider Fig. 25.44 (a) with electrode 1 as a current source, electrode 2 as a current sink, and the four transverse electrodes 3, 4, 5, and 6 as voltage probes. Consider the 2DEG in the Hall bar to be in the $\nu = 2$ quantum Hall state, when there are two edge states. Because of the presence of, and the nature of the voltage probes, the two edge modes that are injected from electrode 1 "bounce off" electrodes 3 and 5 without losing energy, and bring these electrodes to the same potential as electrode 1. Thus $V_5 = V_3 = V_1$, and the current flowing between them is exactly $(2q^2/h)(V_1 - V_2)$, where $V_2 = 0$. Similarly electrodes 2, 4, and 6 are at the same potential (e.g. grounded). There are also localized loops of current flowing far from the edges, but they do not contribute to the externally measured conductances.

Fig. 25.44 (b) shows that when the carrier density is lowered so that the Fermi level enters into the middle of the Landau band, there are extended states formed by the physical proximity of the inner edge state on one side to states from the other side. Tunneling between these states leads to back-scattering and therefore a transition to a non-quantized σ_{xy} and a non-zero σ_{xx} occurs.

At a slightly lower density, the inner edge state vanishes as shown in Fig. 25.44 (c) by breaking up into closed loop paths in the bulk, or paths that start and end in the electrodes. At this density the σ_{xy} and σ_{xx} values approach its $\nu = 1$ value, since the outermost edge mode has survived. Further decrease of the carrier concentration puts the Fermi level between the $n = 1$ and $n = 2$ Landau level bottom parts of the bands, and the system as shown in Fig. 25.44 (d) is in the $\nu = 1$ condition completely, with $\sigma_{xy} = q^2/h$ and $\sigma_{xx} = 0$ again.

The discussion above is for a varying 2DEG density at a constant magnetic field. This is how the integer quantum Hall effect was experimentally discovered. In many experiments, the magnetic field is changed and the 2DEG density is kept fixed. In that case, the Landau level separation and degeneracy changes with B, as indicated schematically in Fig. 25.45, leading to the same arguments as have been discussed for the formation of the steps and zeros characteristic of the quantized Hall effect.

The above model of the quantum Hall effect, that takes careful account of the role of the electrodes, was proposed by Buttiker. It is the most closely connected to what is actually measured and observed in experiments. It has been extended to explain the fractional quantum Hall effect. Importantly, several predictions from this model have been experimentally confirmed. The Landauer–Buttiker formalism also explains conductance quantization in the absence of magnetic fields (Chapter 17, Section 17.2), and forms the basis of the modern understanding of most nanoscale transport phenomena in semiconductors and metals.

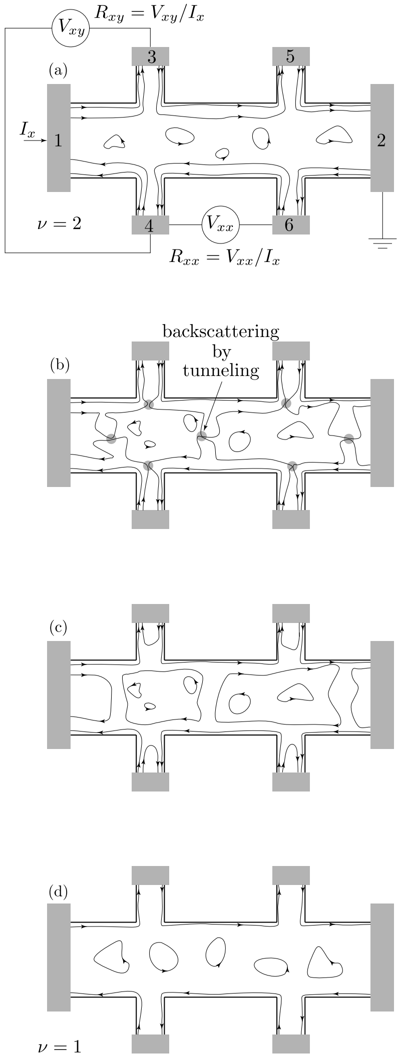


Fig. 25.44 Illustration of the effect of decreasing n_{2d}, or increasing B. (a) Center of orbit motion between a Hall-bar electrodes along edge states at $\nu = 2$, and around localization centers. (b) When the B field is increased the inner edge states from opposite edges come close at some points because of the disorder potential where tunneling to the opposite edge is possible. (c) At yet higher B, backscattering due to tunneling breaks up an edge state in the $\nu = 2 \rightarrow \nu = 1$ transition. (d) Edge states in the $\nu = 1$ state.

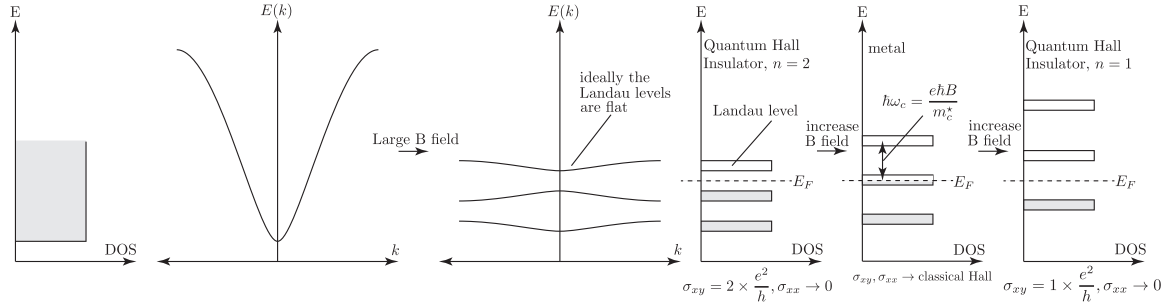

E
DOS
$E(k)$
k
Large B field
$E(k)$
ideally the Landau levels are flat
k
E
Quantum Hall Insulator, $n = 2$
Landau level
E_F
DOS
$\sigma_{xy} = 2 \times \frac{e^2}{h}, \sigma_{xx} \to 0$
increase B field
E
metal
$\hbar\omega_c = \frac{e\hbar B}{m_c^\star}$
E_F
DOS
$\sigma_{xy}, \sigma_{xx} \to$ classical Hall
increase B field
E
Quantum Hall Insulator, $n = 1$
E_F
DOS
$\sigma_{xy} = 1 \times \frac{e^2}{h}, \sigma_{xx} \to 0$

Fig. 25.45 The quantum Hall effect due to changing DOS as a function of the magnetic field.

Several other theoretical models have been proposed for explaining the quantum Hall effects. Among them, R. Laughlin's theory based on gauge invariance, and of a macroscopic wavefunction was one of the first. Jain's composite fermion theory was described earlier, which helped demystify fractional steps. The composite fermion theory has been generalized and extended into a Chern–Simon field theory. The reader can explore these in the references at the end of the chapter.

Fig. 25.46 David Thouless and collaborators discovered, via the Berry phase and Kubo formalism approach, a connection of the quantized Hall conductance to the Chern number in topology. Interestingly this connection was made before M. Berry's discovery of the missing phase in quantum mechanics! Thouless was awarded the Nobel Prize in Physics in 2016 for his contributions to topological phase transitions and topological phases of matter.

To finish the exploration of models, we discuss a mathematically intriguing theory that proposes a connection of the precision of the quantized Hall conductances to the existence of **topological quantum numbers**, pioneered by David Thouless (Fig. 25.46). We will only briefly outline the idea here, and refer the reader to references at the end of the chapter to explore further. As was discussed in Chapter 9, the Berry curvature leads to an additional component to the group velocity of Bloch states. Consider then the classical Newton's law

$$\frac{d\mathbf{p}}{dt} = -e\nabla_r V + e\left(\frac{d\mathbf{x}}{dt} \times \mathbf{B}\right), \tag{25.55}$$

where $\mathbf{p}$ is the momentum of a particle, V is the electrical potential, and $\mathbf{B}$ is the magnetic field. This simply states that the rate of change of momentum of an electron wavepacket is the Lorentz force. The dual of this equation in quantum mechanics because of $[x, p] = i\hbar$ is

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \nabla_p E(k) + \frac{1}{\hbar}\frac{d\mathbf{p}}{dt} \times \mathbf{\Omega} \\ \implies \frac{d\mathbf{x}}{dt} &= \frac{1}{\hbar}\nabla_k E(k) + \frac{d\mathbf{k}}{dt} \times \mathbf{\Omega} = \frac{1}{\hbar}\nabla_k E(k) + \frac{q}{\hbar}\mathbf{E} \times \mathbf{\Omega}, \end{aligned} \tag{25.56}$$

where $\mathbf{\Omega}$ is the Berry curvature. Note the dimensions here. The velocity is in m/s, so the anomalous velocity term is also in m/s. The units of $\mathbf{\Omega} = \nabla_k \times \mathcal{A} = \nabla_k \times i\langle n|\nabla_k|n\rangle =\text{m}^2$, and therefore the anomalous velocity unit is $\frac{1}{\text{m.s}} \cdot \text{m}^2 = \text{m/s}$. The Berry curvature is obtained from the Bloch functions $u_n e^{i\mathbf{k}\cdot\mathbf{r}}$ using the prescription

$$\mathbf{v}_n(\mathbf{k}) = \frac{1}{\hbar}\frac{\partial E_n}{\partial \mathbf{k}} - i\Big(\langle\frac{\partial u_n}{\partial \mathbf{k}}|\frac{\partial u_n}{\partial t}\rangle - \langle\frac{\partial u_n}{\partial t}|\frac{\partial u_n}{\partial \mathbf{k}}\rangle\Big)$$

$$\implies \mathbf{j} = \frac{1}{V}\sum_{n\mathbf{k}} -e\mathbf{v}_n(\mathbf{k}) = \frac{e^2}{\hbar}\sum_n \int \frac{d^2\mathbf{k}}{(2\pi)^2}\boldsymbol{\Omega}_n(\mathbf{k}) \times \mathbf{E}$$

$$\implies \sigma_{xy} = \frac{j}{E} = \frac{e^2}{h}\sum_n \frac{1}{2\pi}\int d^2\mathbf{k}\Omega_n(\mathbf{k}) \implies \boxed{\sigma_{xy} = \frac{e^2}{h} \times C.} \quad (25.57)$$

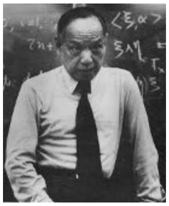


Fig. 25.47 Shiing-shen Chern was a mathematician whose contributions to topology are central to our understanding of the quantum Hall effect. The association of the quantum Hall conductivity with the topological quantum number called Chern number has been proposed as a reason for its high precision.

The integral of the Berry curvature $\boldsymbol{\Omega}_n$ over (k_x, k_y) in the BZ for the n^{th} Landau band is constrained by a mathematical theorem due to Chern (Fig. 25.47) to be $2\pi \times$ (an integer C). This theorem is robust to the presence of disorder in the crystal. The connection to the Chern number can be extended to include the fractional quantum Hall effect. The Chern number connection actually predicts that σ_{xy} should be quantized to integer values not only in 2D systems but in 3D systems as well. In fact this is experimentally confirmed in superlattice structures. While ρ_{xy} in 3D does exhibit plateaus as in 2D, ρ_{xx} does not go to zero. This connection to topology also has led to the prediction and experimental confirmation of several new Hall effects which are described briefly in the next section.

25.12 Hierarchy of Hall effects

The classical and quantum Hall effect is observed for both electrons and holes in semiconductor heterostructures due to the interaction of electron **charge** with magnetic field. For this section we call them the charge Hall effects. There are several closely related Hall effects that occur due to spins and/or valleys in non-magnetic semiconductors. In magnetic materials, an "anomalous" Hall effect is observed due to internal magnetization. Just like the quantization of the standard charge Hall effect, materials that exhibit the quantum spin Hall effect and quantum anomalous Hall effect have been identified. This area of exploration has led to the discovery of a class of materials called topological insulators, which has helped classify electronic states of matter based on the mapping of their bandstructure to different mathematical topological classes. In this section, we give a short, experimentally motivated discussion of these developments, and refer the reader to the references at the end of the chapter for further exploration.

The various types of Hall effects are schematically shown in Fig. 25.48. As we have discussed in this chapter, the Hall effect is observed for all temperatures and all magnetic fields in semiconductors. The quantum magnetotransport effects such as Shubnikov de Haas oscillations and the quantum Hall effect are observable at low temperatures and high magnetic fields when $k_b T << \hbar\omega_c$, below the shaded line shown in Fig. 25.48. As we discussed in detail earlier in this chapter, an additional requirement on the purity of the samples to observe

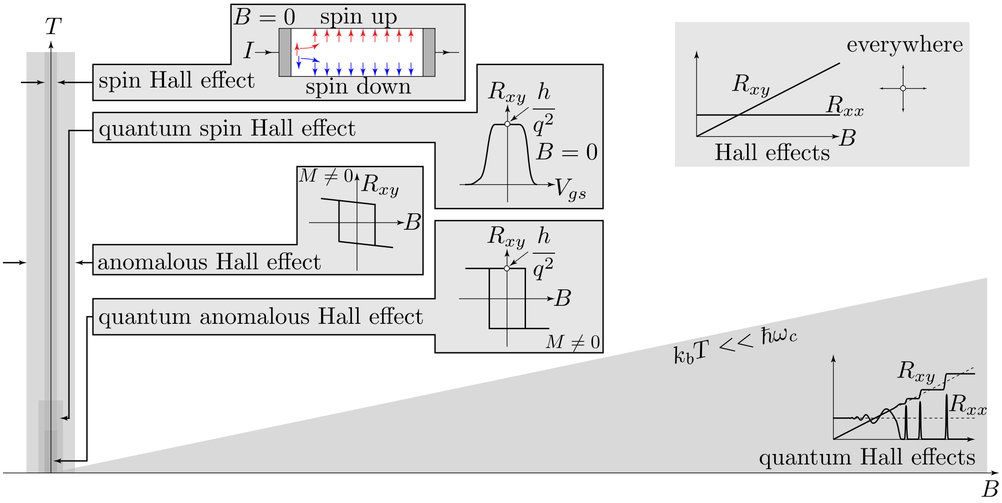

T
B = 0 spin up
I
spin Hall effect
spin down
quantum spin Hall effect
R_{xy} $\frac{h}{q^2}$
B = 0
V_{gs}
M ≠ 0 R_{xy}
B
anomalous Hall effect
R_{xy} $\frac{h}{q^2}$
B
quantum anomalous Hall effect
M ≠ 0
everywhere
R_{xy}
R_{xx}
B
Hall effects
$k_bT << \hbar\omega_c$
R_{xy}
R_{xx}
quantum Hall effects
B

Fig. 25.48 A variety of Hall effects shown schematically in a temperature-magnetic field plot.

Shubnikov de Haas and quantum Hall effects is $\omega_c\tau >> 1$, which is not shown in this figure. Thus, what is also not implicit in this figure is that different semiconductor materials will have different crossover points depending on their carrier effective masses, bandstructures, and purity. Similarly, the relative importance of the various Hall effects, discussed next, strongly depends on the type of the material: (e.g. semiconductors have a strong Hall effect, but metals have a weak Hall effect).

We now focus on the other Hall effects. A general expression for the Hall resistivity written in terms of the external magnetization field H takes the form

$$\rho_{xy} = R_H H + R_{ISHE}(H)H + R_{VHE}(H)H + R_{AHE}M(H) + ... \quad (25.58)$$

where the first term gives the Hall coefficient, the second the inverse spin Hall effect, the third the valley Hall effect, and the fourth term the anomalous Hall effect in terms of the magnetization **M** of the material. This list can be expanded with other effects. Specific properties of the material bandstructures and Fermi level control can lead to precise quantization of the respective Hall resistivities, just as we discussed for the standard (charge-) Hall effect earlier in this chapter.

The spin Hall effect: As shown in Fig. 25.48, when a charge current is injected into semiconductors that possess a strong spin-orbit coupling, the spin-up carriers experience a transverse force in one direction, and spin-down carriers in the opposite direction. This causes an accumulation of spin-up carriers on one edge and of spin-down carriers on the other edge of the sample. The spin Hall effect was predicted by Dyakonov and Perel in 1971 purely based on symmetries, and observed experimentally in the early 2000s. Though this does not create a charge-potential difference in the transverse direction, it does create a spin-potential difference. This means that voltage probes that are sensitive to the spin-polarization (e.g. magnetic contacts) can the detect a spin current. The spin Hall effect is strong in various heavy metals. The inverse spin Hall effect is used to detect and write data in magnetic memories.

Quantum spin Hall effect: Some semiconductors with large spin-orbit coupling undergo band-inversion and develop Dirac-cones. When their Fermi level is tuned to the Dirac point the spin Hall effect is quantized. This makes it possible to observe quantization of the Hall resistance in the $B = 0$ limit, as indicated in Fig. 25.48. Borrowing the idea of edge states of the charge-quantum Hall effects, the quantum spin Hall effect also is believed to result in edge currents that lead to quantization. This sort of quantization is experimentally not observed to be as precise as the charge quantum Hall steps to date. Also this form of quantization has similarities to the ordinary conductance quantization in ballistic conductors, and typically care must be exercised to distinguish the effects. The quantum spin Hall effect has been observed in HgTe/HgCdTe quantum wells, where the quantized conductance is believed to flow at the edges of the 2DEG. Building on this idea, several models of bandstructure have been proposed that exhibit topologically non-trivial properties ranging from 2D to 3D **topological insulators** to **Weyl semimetals**. The purported bandstructures can exhibit a large range of remarkable transport phenomena, but current efforts are primarily wrestling with attaining the crystalline and chemical purity needed to unmask these effects.

The valley Hall effect: Similar to the spin-polarization of each electronic state, several semiconductors exhibit multiple degenerate valleys. Effective valley polarization of carriers at the opposite edges of a Hall bar is achieved at zero magnetic field, due to an electric field alone. Since valley polarization sensitive electronic contacts do not exist to date, this effect is only detectable in polarized photoluminescence measurements. Since intervalley scattering is typically more severe than spin scattering, the effect is difficult to measure[38].

Anomalous Hall effect: Right after Edwin Hall's discovery of the ordinary Hall effect, the method was applied to magnetic materials with $M \neq 0$ as indicated in Fig. 25.48. Anomalously high Hall resistances were observed. In the mid 1950s Karplus and Luttinger explained this phenomenon via the Berry-phase term contribution to the anomalous velocity. Crudely speaking, the magnetization produces a large internal magnetic field, producing a Hall resistance in the absence of an external magnetic field. The Hall resistance R_{xy} exhibits a roughly square hysteretic loop with a characteristic coercive field for a material. This loop is similar to the magnetization loop of the material. In addition to magnetic metals, this effect is also observed in conductive magnetic semiconductors such as GaMnAs.

Quantum anomalous Hall effect: For very special materials that are in a sense hybrids of the conflicting material families: topological

[38] It is possible to use polarized light to selectively excite and detect emission from individual valleys in a semiconductor.

Table 25.2 Hierarchy of Hall effects

Hall Effect	Property	Materials	Temperature	Fields	Robustness to disorder	Topological protection?
Ordinary Hall Effect	Voltage difference transverse to a charge current and magnetic field	Semiconductors, Semimetals, Metals	all	**E**, **B**	yes	no
Quantum Hall Effect	Hall resistance plateaus at integer or special fraction times q^2/h	2D carriers (Si, GaAs, GaN, ZnO, pure semiconductors, Graphene)	low	high **B**, small **E**	Disorder necessary but not high	yes
Spin Hall Effect	spin of opposite type accumulate perpendicular to charge current	semiconductors with spin-orbit interaction (e.g., GaAs)	all	only **E**	yes	no
Quantum Spin Hall Effect	quantized Hall resistance at zero external **B**	2D carriers in heavy atom semiconductors	low	only **E**	perhaps	expected
Anomalous Hall Effect	finite Hall voltage in a magnetic material at zero external **B**	magnetic conductors and semiconductors	all	only **E**	no	no
Quantum Anomalous Hall Effect	quantized Hall resistance in a magnetic material at zero external **B**	special magnetic conductors	very low	only **E**	yes	yes

insulators and ferromagnetic metals, the anomalous Hall resistance R_{xy} is found to be quantized at zero external magnetic fields. Fig. 25.48 shows this behavior mimicking the quantum Hall effect at $B = 0$. The materials palette in which this effect is observed is currently small, but expected to expand in the future.

Table 25.2 summarizes the several Hall effects. Note that this does not include the thermal effects that were discussed in the semi-classical situation in the early sections of this chapter. It is safe to say that there is still much to be discovered when it comes to the rich range of phenomena connected to the idea of the Hall effect!

25.13 Chapter summary section

In this chapter, we learned:

- The classical Hall effect in infinite and finite conductors,
- The effect of magnetic field on magnetotransport properties in semiconductors,
- The combined effects of thermal gradients with magnetic fields (Nernst and Ettingshausen effects),
- The quantum mechanical effect of magnetic fields on the energy levels of atoms, the Zeeman effect and the distinction between orbital and spin magnetization,
- The magnetic quantum oscillations such as the de Haas van Alphen effect and the Shubnikov de Haas effect, and how to use them to extract several microscopic properties of semiconductors,
- The integer and fractional quantum Hall effects in 2DEGs, and several theories that explain the reason for the σ_{xy} plateaus and concomitant vanishing σ_{xx}, and the reasons for the precision of quantization, and
- The other Hall effects such as spin-, valley-, and anomalous-Hall effects, and their quantization.

Further reading

Semiconductor Physics by K. Seeger treats the effects of magnetic fields on Boltzmann transport in significant detail, and also has informative chapters on quantum magnetotransport effects of Shubnikov de Haas oscillations and the quantum Hall effects. **Solid State Physics** by G. Grosso and G. P. Parravicini has excellent chapters on magnetic effects in solid state, magnetic susceptibilities, and the quantum Hall effects. Chapter 6 on Magnetic Quantum Effects written by L. M. Roth and P. N. Argyres in the series **Semiconductors and Semimetals, Vol.1** (1966) edited by R. K. Willardson and A. C. Beer is the standard reference for Shubnikov de Haas oscillations and related phenomena in semiconductors. **Magnetic Oscillations in Metals** by D. Shoenberg is a classic and highly recommended reading for researchers in this area since it takes you through the historical development and all progress till late 1980s. **Semiconductor Nanostructures** by T. Ihn has chapters dedicated to magnetotransport and quantum effects in semiconductors. The last two chapters of **Physics of Interacting Electrons in Disordered Systems** by H. Kamimura and H. Aoki discusses the integer quantum Hall effect in well illustrated detail, and the pre-composite fermion theories of the fractional quantum Hall effect. Several books have been written on the theoretical aspects of the quantum Hall effects such as **The Quantum Hall Effect** by D. Yoshioka, **The Quantum Hall Effects** by T. Chakraborty and P. Peitiläinen, and **composite fermions** by J. Jain. In Chapter 4 on The Quantum Hall Effect in Open Conductors in the series **Semiconductors and Semimetals, Vol. 35** (1992) edited by M. Reed, M. Büttiker provides one of the clearest expositions of the impact of contacts and open systems on the quantum Hall effect with a rich description of experimental details and their connection to the Landauer-Büttiker transmission model of coherent transport. An equivalent exposition written by experimentalists would be highly welcome in this field! **A Short Course on Topological Insulators** by J. K. Asbóth, L. Oroszlány, and A. Pályi is a short and sweet introduction to topological aspects in condensed matter physics and is highly recommended to develop a systematic understanding of the subject from the basics.

Exercises

(25.1) **History of the Hall effect**
The history of the discovery of the Hall effect is rather fascinating in itself because it resulted from an error in Maxwell's judgement of the effect of a magnetic field on a current carrying conductor. In Edwin Hall's 1879 paper ["On a New Action of the Magnet on Electric Currents", American Journal of Mathematics, Volume 2, Page 287-292 (1879).], he writes about the confusion in Maxwell's prediction and his experimental design to test the effect. The report is written candidly and should be read to appreciate the spirit of scientific enquiry, difficulties encountered in experimental work, and the value of persistence.

(a) Read Edwin Hall's original paper on the Hall effect.

(b) Describe why Maxwell's prediction that a magnetic field should not have an effect on the conductor, but *not* the "current carriers" was confusing for Edwin Hall.

(c) Describe the interactions described between Hall and his supervisor Prof. Rowland, and the role played by his supervisor in this discovery.

(d) Describe how the experiment was designed to test the hypothesis that Maxwell was incorrect.

(e) Describe the difficulties in the experiment, and how it finally revealed a strong Hall voltage. What was the trick?

(f) Write how even when "wrong", Maxwell inspired excellent research directions of high value.

(25.2) **Hall factors in generalized bandstructures and dimensions**
Prove by using partial integration the generalized expressions derived for the Hall factor in Equation 25.13. Explore how the Hall factor changes as the dimensionality is lowered.

(25.3) **Magnetoresistance for non-parabolic bandstructure**
In this chapter we discussed exclusively the magnetotransport properties of semiconductors with parabolic bandstructures. The results are easily generalized to non-parabolic bandstructures. Answer the following specifically for a conical bandstructure:

(a) Calculate the Hall factor r_H for the conical bandstructure $E(k) = \hbar v_F |\mathbf{k}|$ for d dimensions. In the specific case of $d = 2$ the bandstructure represents one valley of monolayer graphene. Compare and contrast with the results for parabolic bandstructure.

(b) Discuss the changes to the theory of Shubnikov de Haas oscillations in graphene compared to a semiconductor with parabolic bandstructure.

(c) Calculate the Landau levels of 2D graphene and show that unlike for parabolic bandstructure, the Landau level separation in energy is not uniform in energy. Specifically, because the bandstructure of 2D graphene $E(k) = \pm\hbar v_F k$ has a positive and negative branch, show that there is one Landau level formed at energy $E = 0$.

(25.4) **Commutation relations of the velocity in a magnetic field**
Consider the velocity operator $\hat{\mathbf{v}} = \frac{-i\hbar\nabla + q\mathbf{A}}{m_e}$ for electrons of charge $-q$ and mass m_e in the presence of a magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$.

(a) Show that $[\hat{v}_x, \hat{v}_y] = -\frac{iq\hbar}{m_e^2} B_z$, and the corresponding cyclic relations for the velocity components. In other words, the Cartesian coordinates of the velocity vector do not commute in the presence of a magnetic field. This should make intuitive sense, since the magnetic field curves the motion of electrons perpendicular to it.

(b) Combining the above, show that the magnetic field may be written in the form $\mu_B \mathbf{B} = \frac{i}{2} m_e([\hat{v}_y, \hat{v}_z]\hat{x} + [\hat{v}_z, \hat{v}_x]\hat{y} + [\hat{v}_x, \hat{v}_y]\hat{z})$, where $\mu_B = q\hbar/2m_e$ is the Bohr magneton. This is written more compactly as $\frac{i}{2} m_e \hat{\mathbf{v}} \times \hat{\mathbf{v}} = \mu_B \mathbf{B}$. Note that both sides have units of energy.

(25.5) **Quantum mechanical currents in a magnetic field**
Show using Exercise 25.4 that in the presence of a magnetic field, the quantum current density due to

electrons of charge $-q$ and mass m_e in a wavefunction state ψ can be expressed as a sum of an orbital and spin components $\mathbf{J} = \mathbf{J}_{\text{orbital}} + \mathbf{J}_{\text{spin}}$, where

$$\mathbf{J}_{\text{orbital}} = -\frac{iq\hbar}{2m_e}(\psi\nabla\psi^\star - \psi^\star\nabla\psi) - \frac{q^2}{m_e}\psi^\star\psi\mathbf{A}, \tag{25.59}$$

and a spin component of the current

$$\mathbf{J}_{\text{spin}} = \mu\nabla \times [\psi^\star \frac{\hat{\mathbf{s}}}{s}\psi] \tag{25.60}$$

where μ is the permeability, $\mathbf{s}$ is the spin, and $s = |\mathbf{s}|$ is its magnitude.

(25.6) **Landau levels from the Schrödinger equation**
(a) Prove that the solution of the Schrödinger equation with a magnetic vector potential is indeed given by Equation 25.30.

(b) Prove that the solution of eigenvalues are identical for other gauge choices, but the eigenfunctions can be different. Discuss the concept of gauge invariance and why the different eigenfunctions can still describe the same physics and observables.

(25.7) **Poisson summation formula**
(a) Explain the Poisson summation formula in Equation 25.41. Read about its importance in mathematics and physics.

(b) Prove Equation 25.42.

(c) From part (b), using Equations 25.43 prove Equation 25.44. Take the derivative of the second term of Equation 25.44 to prove the Landau steady diamagnetism.

(25.8) **Pauli paramagnetism vs. Landau diamagnetism**
(a) Prove the results of Pauli paramagnetism in Equation 25.37 which is for electron motion in 3D.

(b) Argue why this result is independent of the effective mass of electrons.

(c) Discuss how this result would change in lower dimensions.

(d) Compare and contrast the Pauli paramagnetic susceptibility with Landau diamagnetism. Show why for free–electrons $\chi_m(\text{dia.}) = -\frac{1}{3}\chi_m(\text{para.})$.

(e) Show that the order of magnitudes of the susceptibilities for free–electrons are: for Pauli paramagnetism $\chi_m \sim 10^{-4}$ and for Landau diamagnetism $\chi_m \sim -10^{-6}$.

(f) Discuss why the Landau diamagnetic susceptibility depends on the effective mass $m^\star$ of electrons whereas the Pauli paramagnetism does not. Argue why this explains the experimental fact that some metals exhibit paramagnetism whereas some others exhibit diamagnetism.

(g) Show that the crossover effective mass for Landau diamagnetism to dominate is $m^\star \leq \frac{m_e}{\sqrt{3}}$.

(25.9) **Quantum Hall effects**
Consider Fig. 25.33 for this question.
(a) Using the given 2DEG density, draw the classical Hall resistance R_{xy} line and confirm that the quantum Hall plateaus indeed cross this line such as indicated schematically in Fig. 25.32 (c).

(b) Verify the filling factors indicated, and comment on the fractional step. Why is the fractional step not clearly resolved?

(25.10) **Integer quantum Hall effect**
Consider Fig. 25.34 for this question.

(a) Given the magnetic field, calculate the carrier concentration vs. the gate voltage and therefore the gate capacitance per unit area from the integer quantum Hall steps.

(b) Argue why the shape of R_{xx} must be the derivative of R_{xy}.

(c) What determines the width of the plateaus of R_{xy} and the corresponding zeroes of R_{xx}?

(25.11) **Magnetocapacitance oscillations**
Consider a gated 2DEG in a semiconductor field-effect transistor subjected to a large magnetic field at low temperatures. Show that the 2DEG carrier concentration changes with the Fermi level as

$$\frac{dn_{2d}}{dE_F} = \frac{m^\star}{\pi\hbar^2}[1+2\sum_{r=1}^{\infty}(-1)^r e^{-\frac{\pi}{\omega_c\tau_q}r}\frac{\chi_r}{\sinh\chi_r}\cos(2\pi r\frac{E_F}{\hbar\omega_c})], \tag{25.61}$$

where $\chi_r = 2\pi^2 r k_b T/(\hbar\omega_c)$ and all symbols have their usual meanings. From this relation, plot how the gate capacitance changes with the gate voltage. Experimentally, the C_{gs} vs. V_{gs} plot exhibits magnetocapacitance oscillations. Discuss the analogy of magnetocapacitance oscillations to the de Haas van Alphen oscillations in magnetization.

(25.12) **Quantum Hall effects and quantized conductance: similarities and differences**
The quantization of conductance was discussed in Chapter 17. Fig. 17.4 showed that the conduc-

tance went through quantized steps as a function of the gate voltage: this was first observed in 1988. Fig. 25.34 shows the integer quantum Hall effect, in which R_{xy} goes through quantized steps as a function of the gate voltage. This phenomena was first observed in 1980. They are different phenomena, in spite of their appearance. Explain how are they different, and the connection between the quantized conductances in each case.

(25.13) **Spin-orbit coupling in 1D and Majorana fermions**

The Dirac equation had predicted the existence of antimatter for the electron: the positron. The positron has the same mass, but opposite charge as the electron. When matter and antimatter meet, they annihilate each other, emitting light. If an electron meets a positron, they annihilate each other, releasing a photon of energy $h\nu = 2m_e c^2$. A few years after Dirac's prediction, the positron was experimentally discovered, and is today used in several experiments.

Now roughly a decade after Dirac, Ettore Majorana discovered another equation that predicted a strange particle – this particle was its own antiparticle! What this means is this particle cannot be annihilated and should have higher stability than those that have antiparticles. The basic ingredients for these particles, called **Marjorana fermions**, is that they are **chargeless**, and also **spinless**. As a result, they do not respond to electric fields, or magnetic fields! This state is one of the most sought-after electronic phases in many laboratories today.

For this problem, we do not need to worry much about the exact properties of the Majorana Fermion, but the arguments to produce one out of a semiconductor are straightforward. In the preceding chapters, we have covered most of the necessary ingredients to understand this proposal. In addition to the idea of an effective mass in the conduction band, we add three ingredients to make a standard 1D-fermion state in a crystal into a Majorana state.

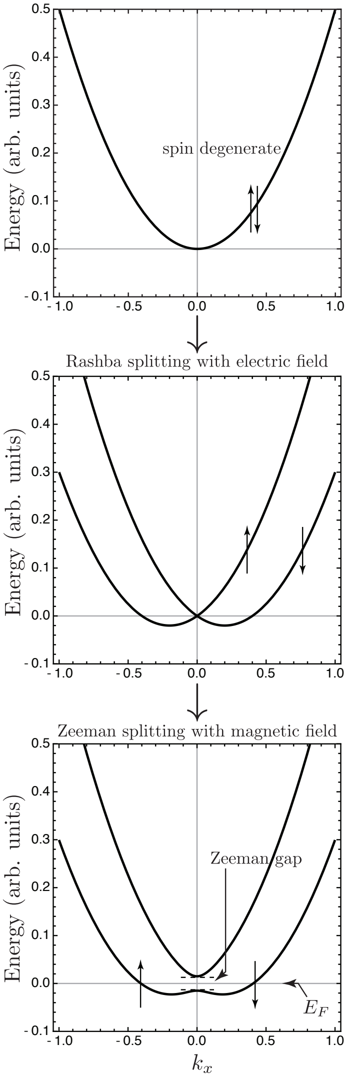


Fig. 25.49 Evolution from a standard 1D nanowire electronic states to a Majorana Fermion state.

The effective Hamiltonian for an electron in the presence of a magnetic field and a strong spin-orbit coupling is

$$H = \underbrace{\frac{\hbar^2 k^2}{2m_c^\star}\sigma_0}_{\text{band}} + \underbrace{\frac{g^\star \mu_B}{2}(\vec{\sigma}\cdot\mathbf{B})}_{\text{Zeeman}} + \underbrace{\alpha_R(\vec{\sigma}\cdot\mathbf{k}\times\mathbf{E})}_{\text{Rashba}}, \tag{25.62}$$

where the σ_0 and $\sigma = (\sigma_x, \sigma_y, \sigma_z)$ are the Pauli spin matrices as described in Exercise 6.3, **B** is a mag-

netic field, and $\mathbf{E}$ is an electric field. $\mathbf{k}$ is the electron wavevector, $g^\star$ is an effective Lande Zeeman coupling term, and $\mu_B = \frac{q\hbar}{2m_e}$, the Bohr-Magneton is a constant, as is α_R, called the Rashba coefficient.
(a) We will solve a more general problem first. Consider the conduction band of a 2D crystal with allowed wavevectors $\mathbf{k} = k_x\hat{x} + k_y\hat{y}$, an electric field $\mathbf{E} = E_z\hat{z}$, and a magnetic field $\mathbf{B} = B_z\hat{z}$, both perpendicular to the 2D plane. Show that the Hamiltonian is

$$H = \begin{bmatrix} \frac{\hbar^2 k^2}{2m_c^\star} + \frac{g^\star \mu_B}{2} B_z & \alpha_R E_z(k_y + ik_x) \\ \alpha_R E_z(k_y - ik_x) & \frac{\hbar^2 k^2}{2m_c^\star} - \frac{g^\star \mu_B}{2} B_z \end{bmatrix}. \quad (25.63)$$

(b) We will come back to how the magnetic field and electric field can be used to cause spin-polarization in 2D electron systems. For this problem, let us reduce the 2D system to 1D aligned along the x-axis. This may be achieved with a nanowire aligned along the x-axis. Write the Hamiltonian for $k_y = 0$, find the eigenvalues, and show that the result for $E_z = 0$ and $B_z = 0$ is the bandstructure in Fig. 25.49 (a). This is simply a standard 1D conductor in which each k-state has a spin-degeneracy of $g_s = 2$.

(c) Now some materials boast a physical mechanism called a strong spin-orbit coupling. This is an internally generated magnetic field that is present even in the absence of an external $\mathbf{B}$ field. The origin is in structural inversion asymmetry (SIA) of the underlying crystal. The Rashba effect term in Equation 25.62 arises because of this effect. By applying a non-zero electric field E_z (and zero external magnetic field B_z) to the 1D electron system, show that the spin-degenerate conduction band is split along the k_x axis into two bands of different spins, as shown in Fig. 25.49 (b).

(d) Note that the $k_x = 0$ state is still spin-degenerate. Explain why this spin splitting is only achieved due to motion of the electron. This is the origin of the name "spin-orbit" term. Explain why no matter where is the Fermi level in this system, there are always four spins at that energy, two of one spin, and two of the other. The last remaining spin-degeneracy at $k_x = 0$ could be removed if we could open a gap at $k_x = 0$.

(e) Separate from the Rashba effect is the ordinary Zeeman effect of spin splitting of degenerate states in the presence of an external magnetic field, given by the second term of Equation 25.62. The Zeeman splitting does not depend on the electron k, and equal for all k states. Verify this is the case, and the total splitting is $g^\star \mu_B B_z$.

(f) Finally, if the Zeeman and the Rashba effects are turned on simultaneously, show that the bandstructure becomes as shown in Fig. 25.49(c). If now the Fermi level is placed as shown, the right-going electron has up spin, and the left-going electron has down spin. So the electrons have exactly opposite momenta, and exactly opposite spins. Note that this is a rather exotic material that will have one type of spin if current flows to the right, and other type of spin for current flowing to the left.

(g) In this book, we do not cover the last remaining ingredient to make the Majorana fermion. If the two states at the Fermi level in Fig. 25.49 could be strongly paired or glued to form a composite object, it will be spinless. This pairing is achieved by placing the nanowire on a *superconductor*. In addition to pairing the electrons of opposite momentum and opposite spin (called a spin-singlet Cooper pair), the superconducting effect also makes them effectively charge-less, by opening a small electron/hole gap at the Fermi level. But that is beyond what this book intends to cover!

Part IV

Quantum Photonics with Semiconductors

Let There Be Light: Maxwell Equations

26

In this chapter we learn:

- What is the physical content of light and how can it be explained quantitatively?
- How do we understand light in its purest indivisible form via its wavelength or frequency?
- How does light interact with atoms, molecules, and solids, i.e., with material media?

26.1 Maxwell equations in vacuum 704
26.2 Light from Maxwell equations 705
26.3 Maxwell equations in $(\mathbf{k}, \omega)$ space 707
26.4 Maxwell equations in materials 709
26.5 Classical light-matter interaction 711
26.6 Kramers–Kronig relations 712
26.7 Accelerating charges radiate 713
26.8 Need for quantum theory of light 713
26.9 Chapter summary section 714
Further reading 714
Exercises 715

Light has fascinated us for ages. And deservedly so. Everything we know about the earth and the universe is because of light. Light from the sun sustains life on earth. Learning to measure and understand the contents of light has enabled us to understand the origins of the universe in the Big Bang, and to imagine its possible future. And one cannot forget the sheer visual pleasure of a beautiful sunset, a coral reef, or an iridescent flower in full blossom. Indeed, the beauty of light and color is a rare thing that scientists and artists agree to share and appreciate.

Our fascination with light has led to three of the greatest revolutions in 19th and 20th century physics. Sunlight, considered as a ″gift of the Gods″ was thought to be the purest indivisible substance, till Newton observed that passing it through a prism split it into multiple colors. Passing each of the colors through another prism could not split it further. Newton surmised that light was composed of particles (″**corpuscles**″). But in the early 19th century, Young proved that light was a wave because it exhibited interference and diffraction.

Fig. 26.1 James Clerk Maxwell.

Michael Faraday had a strong hunch that light was composed of a mixture of electric and magnetic fields, but could not back up the hunch mathematically. The race for understanding the fabric of light reached a milestone when Maxwell (Fig. 26.1) gave Faraday's hunch a rigorous mathematical grounding. Maxwell's theory combined, in one stroke, electricity, magnetism, and light into an eternal braid[1]. The Maxwell equations constitute a complete **electromagnetic theory** that predict the existence of light as a propagating electromagnetic wave.

The second and third revolutions born out of light occurred in the early 20th century in parallel. Trying to understand the origins of

---

[1] J. R. Pierce famously wrote ″To anyone who is motivated by anything beyond the most narrowly practical, it is worthwhile to understand Maxwell's equations simply for the good of his soul.″

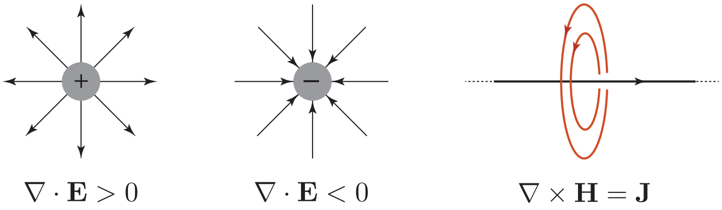

+
−
$\nabla\cdot\mathbf{E}>0$
$\nabla\cdot\mathbf{E}<0$
$\nabla\times\mathbf{H}=\mathbf{J}$

Fig. 26.2 Electrostatic **E** and **H** fields.

blackbody radiation, photoelectric effect, and the spectral lines of hydrogen atoms led to the uncovering of **quantum mechanics**. And Einstein's fascination with the interplay of light and matter, and of space and time led to the theory of **relativity**. Much of modern physics rests on these three pillars of light: that of electromagnetism, quantum mechanics, and relativity. It would be foolhardy to think that we know all there is to know about light. It will continue to amaze us and help probe deeper into the fabric of nature through similar revolutions in the future. In this chapter, we discuss Maxwell's theory of electromagnetism in preparation for the quantum picture, which is covered in the next chapter.

26.1 Maxwell equations in vacuum

Maxwell's equations connect the electric field **E** and the magnetic field intensity **H** to source charges ρ and currents **J** via the four relations[2]

$$\boxed{\begin{array}{lcll}\nabla\cdot\mathbf{D} & = & \rho, & \text{Gauss's law}\\ \nabla\cdot\mathbf{B} & = & 0, & \text{Gauss's law}\\ \nabla\times\mathbf{E} & = & -\frac{\partial\mathbf{B}}{\partial t}, & \text{Faraday's law}\\ \nabla\times\mathbf{H} & = & \mathbf{J}+\frac{\partial\mathbf{D}}{\partial t}, & \text{Ampere's law.}\end{array}} \tag{26.1}$$

Here the source term ρ has units of charge per unit volume (C/m^3), and current source term **J** has units of current per unit area (A/m^2). **H** is related to the magnetic flux density **B** via $\mathbf{B}=\mu_0\mathbf{H}$, and the displacement vector is related to the electric field via $\mathbf{D}=\epsilon_0\mathbf{E}$. The constant ϵ_0 is the permittivity of vacuum, and μ_0 is the permeability of vacuum. They are related by $\epsilon_0\mu_0=1/c^2$, where c is the speed of light in vacuum.

In vacuum, $\mathbf{D}=\epsilon_0\mathbf{E}$, and Gauss's law $\nabla\cdot\mathbf{E}=\rho/\epsilon_0$ says that electric field lines (vectors) due to *static* charges originate at points in space where there are +ve charges, and terminate at negative charges, as indicated in Fig. 26.2. Vectors originating from a point in space have a positive divergence. Gauss's law takes the form of the Poisson equation in semiconductor device physics, and if the charge ρ is zero, it goes by the name of Laplace equation. Gauss's law for magnetic fields tells us that magnetic field lines **B** have no beginnings and no ends: unlike static electric field lines, they close on themselves.

[2] We directly use the results of electromagnetic theory. Readers can review the topics from texts on the subject, some of which are cited at the end of the chapter. We note from the outset that we will exclusively use SI units throughout for Module IV of this book, which focuses on photonics with semiconductors.

Note that in electrostatics and magnetostatics, $\partial(...)/\partial t \to 0$ leads to the static magnetic field relation $\nabla \times \mathbf{H} = \mathbf{J}$. The magnetic field lines curl around a wire carrying a DC current, as shown in Fig. 26.2. Electrostatic phenomena such as electric fields in the presence of static charge such as in pn junctions, transistors, and optical devices in equilibrium, and magnetostatic phenomena such as magnetic fields near wires carrying DC currents are covered by the condition $\partial(...)/\partial t \to 0$. In this case, the electric and magnetic fields are **decoupled**. This means a static charge produces just electric fields and no magnetic fields. A static current (composed of charges moving at a **constant** velocity) produces a magnetic field, but no electric field.

Since in electrostatics, $\nabla \times \mathbf{E} = 0$, the static electric field vector can be expressed as the gradient of a scalar potential $\mathbf{E} = -\nabla\Phi$ because $\nabla \times (\nabla\Phi) = 0$ is an identity. Φ is then the scalar electric potential. However, the same cannot be done for the magnetic field vector even in static conditions, because $\nabla \times \mathbf{H} = \mathbf{J} \neq 0$. Instead, the magnetic field can be written as the **curl** of another *vector* field $\mathbf{A}$ through the relation $\mathbf{B} = \nabla \times \mathbf{A}$, where $\mathbf{A}$ is called the **magnetic vector potential**[3]. Hence from the Maxwell equations, $\mathbf{E} = -d\mathbf{A}/dt$.

Faraday's law says that a time-varying magnetic field creates an electric field. The electric field lines thus produced "curl" around the magnetic field lines. Ampere's law says that a magnetic field intensity $\mathbf{H}$ may be produced not just by a conductor carrying current $\mathbf{J}$, but also by a time-varying electric field in the form of the displacement current $\partial\mathbf{D}/\partial t$. The original Ampere's law did not have the displacement current. Maxwell realized that without it, the four constitutive equations would violate current continuity relations. To illustrate, without the displacement current term, $\nabla \times \mathbf{H} = \mathbf{J}$, and taking the divergence of both sides, we get $\nabla \cdot \nabla \times \mathbf{H} = \nabla \cdot \mathbf{J} = 0$ because the divergence of curl of any vector field is zero. But the continuity equation requires

$$\boxed{\nabla \cdot \mathbf{J} \quad = \quad -\partial\rho/\partial t, \quad \text{Continuity Equation}} \tag{26.2}$$

which is necessary for the conservation of charge. With the introduction of the displacement current term, Maxwell resolved this conflict: $\nabla \cdot \mathbf{J} = -\nabla \cdot \frac{\partial \mathbf{D}}{\partial t} = -\frac{\partial}{\partial t}(\nabla \cdot \mathbf{D}) = -\frac{\partial \rho}{\partial t}$, which connects to Gauss's law.

26.2 Light from Maxwell equations

The displacement current is the crucial link between electricity and magnetism. It explains the nature of light as an electromagnetic wave. Let's first look at this feature qualitatively. Fig. 26.3 shows a metal wire connected to an AC voltage source. The source (or battery) sloshes electrons back and forth from the ground into the wire, causing a charge-density wave as shown schematically. Note that the charge density in the wire is changing *continuously* in time and space. The frequency is ω_0. As a result of charge pileups, electric field lines

---

[3] The vector potential $\mathbf{A}$ plays a central role in incorporating light into the Schrödinger equation to study light-matter interaction. This will be the topic of Chapter 27.

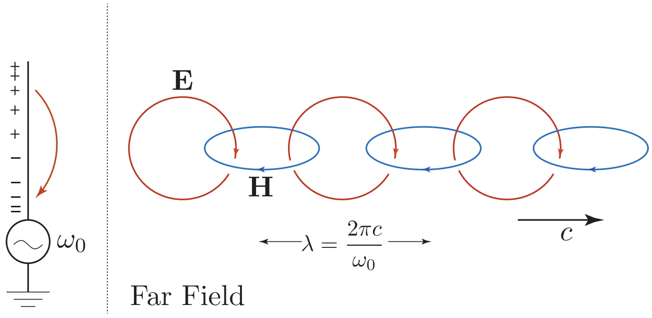


Fig. 26.3 Antenna producing an electromagnetic wave.

emerge from +ve charges and terminate on -ve charges. This electric field is changing in space and time as well, leading to non-zero $\nabla \times \mathbf{E}$ and $\partial \mathbf{E}/\partial t$. The time-varying electric field *creates* a time-varying magnetic field $\mathbf{H}$ because of displacement current. The time-varying magnetic field creates a time-varying electric field by Faraday's law. Far from the antenna, the *fields detach from the source* antenna and become *self-sustaining*: the time-varying $\mathbf{E}$ creates $\mathbf{H}$, and vice versa. An electromagnetic wave is thus born; the oscillations of electric and magnetic fields move at the speed of light c. For an antenna radiating at a frequency ω_0, the wavelength is $\lambda = 2\pi c/\omega_0$. That the wave is self-sustaining is the most fascinating feature of light. If at some time the source was switched off, the far field wave continues to propagate – *forever*, unless it encounters charges again. That of course is how light from the most distant galaxies and supernovae reach our antennas and telescopes, propagating through "light years" in the vacuum of space, sustaining the oscillations[4].

Now let's make this observation mathematically rigorous. Consider a region in space with no charges ($\nabla \cdot \mathbf{D} = \rho = 0 = \nabla \cdot \mathbf{E}$) and no currents $\mathbf{J} = 0$. Take the curl of Faraday's equation to obtain $\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\frac{\partial}{\partial t}(\nabla \times \mathbf{B}) = -\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E}$, where we make use of Ampere's law. Since in a source-free region $\nabla \cdot \mathbf{E} = 0$, we get the *wave equations*

$$\boxed{\begin{aligned} (\nabla^2 - \tfrac{1}{c^2}\tfrac{\partial^2}{\partial t^2})\mathbf{E} &= 0, \quad \text{wave equations} \\ (\nabla^2 - \tfrac{1}{c^2}\tfrac{\partial^2}{\partial t^2})\mathbf{B} &= 0. \end{aligned}} \tag{26.3}$$

Note that the wave equation states that the electric field and magnetic field oscillate both in space and time. The ratio of oscillations in space (captured by ∇^2) and oscillations in time (captured by $\frac{\partial^2}{\partial t^2}$) is the speed at which the wave moves, and it is $c = 1/\sqrt{\mu_0 \epsilon_0}$. The number is exactly equal to the experimentally measured speed of light, which solidifies the connection that light is an electromagnetic wave. We note that just like the solution to Dirac's equation in quantum mechanics is the electron, the solution of Maxwell's wave equation is light (or photons). Thus one can say that light has "emerged" from the solution of Maxwell equations.

[4] Boltzmann wrote "... was it a God who wrote these lines ..." in connection to "Let there be light."

However, we must be cautious in calling the wave equation above representing light *alone*. Consider a generic wave equation $(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2})f(r,t) = 0$. This wave moves at a speed v. We can create a sound wave, *and* also a wave on the surface of water, that move at the same speed v. But their oscillation amplitudes $f(r,t)$ represent distinct physical phenomena[5].

Consider a generic vector field of the type $\mathbf{V}(\mathbf{r},t) = V_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}\hat{\eta}$, where $\hat{\eta}$ is the direction of the vector. This field will satisfy the wave Equations 26.3 if $\omega = c|\mathbf{k}|$, as may be verified by substitution. This requirement is the first constraint on the nature of electromagnetic waves. The second stringent constraint is that the field must satisfy Gauss's laws $\nabla \cdot \mathbf{E} = 0$ and $\nabla \cdot \mathbf{B} = 0$ for free space. In other words, electric and magnetic vector fields are a special class of vector fields. Their special nature is elevated by the physical observation that *no other wave can move at the speed of light*. Einstein's theory of relativity proves that the speed of light in a vacuum is absolute, and unique for electromagnetic waves: every other kind of wave falls short of the speed of light. Thus, Maxwell's wave equation *uniquely* represents light, composed of self-sustaining oscillating $\mathbf{E}$ and $\mathbf{B}$ fields.

26.3 Maxwell equations in $(\mathbf{k}, \omega)$ space

Consider an electromagnetic **plane wave** of a fixed frequency ω. Since $\mathbf{E}, \mathbf{B} \propto e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$, we make two observations. Time derivatives of Faraday and Ampere's laws give $\frac{\partial}{\partial t}e^{-i\omega t} = -i\omega e^{-i\omega t}$, which means we can replace $\frac{\partial}{\partial t} \rightarrow -i\omega$, $\frac{\partial^2}{\partial t^2} \rightarrow (-i\omega)^2$, and so on. Similarly, the vector operators div and curl act on[6] the $e^{i\mathbf{k}\cdot\mathbf{r}}$ part only, giving $\nabla \cdot (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta}) = i\mathbf{k} \cdot (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta})$ and $\nabla \times (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta}) = i\mathbf{k} \times (e^{i\mathbf{k}\cdot\mathbf{r}}\hat{\eta})$. These relations may be verified by straightforward substitution. Thus, we can replace $\nabla \rightarrow i\mathbf{k}$. With these observations, Maxwell Equations 26.1 in free-space become

$$\boxed{\begin{aligned} \mathbf{k} \cdot \mathbf{E} &= 0, \\ \mathbf{k} \cdot \mathbf{B} &= 0, \\ \mathbf{k} \times \mathbf{E} &= \omega \mathbf{B}, \\ \mathbf{k} \times \mathbf{B} &= -\frac{\omega}{c^2}\mathbf{E}. \end{aligned}} \tag{26.4}$$

Note that we have converted Maxwell's equations in real space and time $(\mathbf{r}, t)$ to "Fourier" space $(\mathbf{k}, \omega)$ in this process. Just as in Fourier analysis we decompose a function into its spectral components, light of a particular $\mathbf{k}$ and corresponding frequency $\omega = c|\mathbf{k}|$ is spectrally pure, and forms the "sine" and "cosine" bases. Any mixture of light is a linear combination of these spectrally pure components. For example, white light is composed of multiple wavelengths. Since $\mathbf{B} = \nabla \times \mathbf{A}$, we can write $\mathbf{B} = i\mathbf{k} \times \mathbf{A}$, and hence the magnitudes are related by $B^2 = k^2 A^2 = (\frac{\omega}{c})^2 A^2$. The energy content in a region in space of volume Ω that houses electric and magnetic fields of frequency ω is given by[7]

---

[5] Another analogy is: if a cheetah runs as fast as a car, they are not the same object!

[6] These arguments hold for plane waves, and must be appropriately modified for spherical waves.

[7] Notice a remarkable similarity between the expression for energy of an electromagnetic field with that of a harmonic oscillator energy or Hamiltonian $H_{osc} = \frac{1}{2}m(\frac{dx}{dt})^2 + \frac{1}{2}m\omega^2 x^2$ or that of an LC oscillator $H_{LC} = \frac{1}{2}L(\frac{dQ}{dt})^2 + \frac{Q^2}{2C}$. This is the direct pathway to quantize the electromagnetic field for quantum electrodynamics of light and circuits, as discussed in Chapter 29.

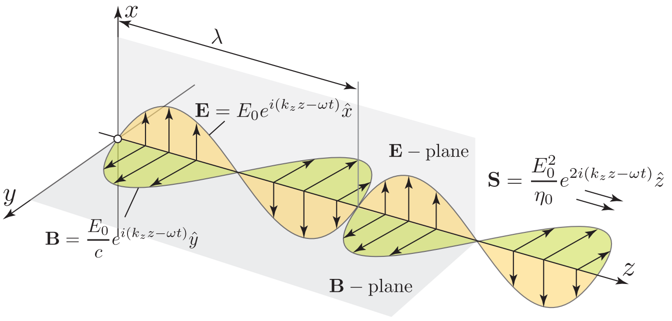

x
λ
$\mathbf{E} = E_0 e^{i(k_z z - \omega t)} \hat{x}$
$\mathbf{E}$ – plane
$\mathbf{S} = \frac{E_0^2}{\eta_0} e^{2i(k_z z - \omega t)} \hat{z}$
y
$\mathbf{B} = \frac{E_0}{c} e^{i(k_z z - \omega t)} \hat{y}$
$\mathbf{B}$ – plane
z

Fig. 26.4 An electromagnetic wave.

$$\boxed{H_{em}(\omega) = \Omega \cdot [\frac{1}{2}\epsilon_0 E^2 + \frac{1}{2}\mu_0 H^2] = \Omega \cdot [\frac{1}{2}\epsilon_0 E^2 + \frac{1}{2}\epsilon_0 \omega^2 A^2].} \quad (26.5)$$

Let us now investigate the properties of a spectrally pure (monochromatic) component of the electromagnetic wave. From Equations 26.4, we note that $\mathbf{k} \perp \mathbf{E} \perp \mathbf{B}$, and the direction of $\mathbf{k}$ is along $\mathbf{E} \times \mathbf{B}$. The simplest possibility is shown in Fig. 26.4. If we align the x-axis along the electric field vector and the y-axis along the magnetic field vector, then the wave propagates along the +ve z-axis, i.e., $\mathbf{k} = k\hat{z}$. The electric field vectors lie in the $x - z$ plane, and may be written as $\mathbf{E}(\mathbf{r}, t) = E_0 e^{i(kz - \omega t)} \hat{x}$, which is a plane wave. For a plane wave, nothing changes in space on the planes perpendicular to the direction of propagation, so the $\mathbf{E}$ field is the same at all $x - y$ planes: $\mathbf{E}(x, y, z) = \mathbf{E}(0, 0, z)$.

From Faraday's law, $\mathbf{B} = \mathbf{k} \times \mathbf{E}/\omega$, and the magnetic field vectors $\mathbf{B}(\mathbf{r}, t) = \frac{E_0}{c} e^{i(kz - \omega t)} \hat{y}$ lie in the $y - z$ plane. Note that here we use $\omega = ck$ and $k = k_z$. The amplitudes of the electric and magnetic fields are thus related by $E_0 = cB_0$, and the relation to magnetic field *amplitude* $\mathbf{H} = \mathbf{B}/\mu_0$ is $E_0 = c\mu_0 H_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} H_0 = \eta_0 H_0$. Since E_0 has units V/m and H_0 has units A/m, η has units of V/A or Ohms. η_0 is called the **wave impedance of free space**; it has a value $\eta_0 \approx 377\ \Omega$.

The direction of propagation of this wave is *always* perpendicular to the electric and magnetic field vectors, given by the right hand rule. Since the field vectors lie on well-defined planes, this type of electromagnetic wave is called *plane-polarized*. In case there was a phase difference between the electric and magnetic fields, the electric and magnetic field vectors will rotate in the $x - y$ planes as the wave propagates, and the wave would then be called circularly or *elliptically* polarized, depending upon the phase difference.

For the monochromatic wave, Maxwell's wave equation becomes $(|\mathbf{k}|^2 - (\frac{\omega}{c})^2)\mathbf{E} = 0$. For non-zero $\mathbf{E}$, $\omega = c|\mathbf{k}| = ck$. The electromagnetic field carries energy in the +ve z-direction. The instantaneous power carried by the wave is given by the Poynting vector $\mathbf{S}(\mathbf{r}, t) = \mathbf{E} \times \mathbf{H} = \frac{E_0^2}{\eta_0} e^{i(kz - \omega t)} \hat{z}$. The units are in Watts/m^2. Typically we are interested in the time-averaged power density, which is given

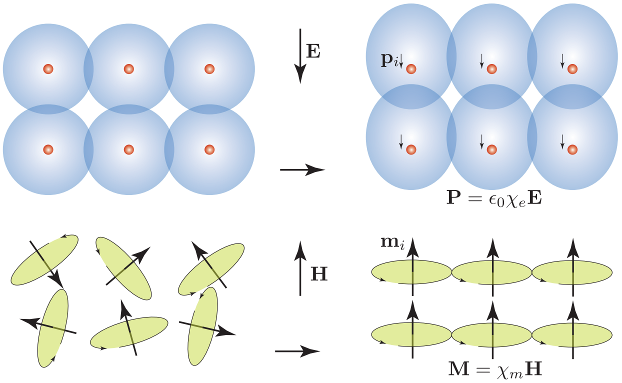

E
$\mathbf{p}_i$
$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}$
H
$\mathbf{m}_i$
$\mathbf{M} = \chi_m \mathbf{H}$

Fig. 26.5 Dielectric and magnetic materials. Orientation of electric and magnetic dipoles by external fields, leading to electric and magnetic susceptibilities.

by

$$\mathbf{S} = \langle \mathbf{S}(\mathbf{r}, t) \rangle = \frac{1}{2}\text{Re}[\mathbf{E} \times \mathbf{H}^\star] = \frac{E_0^2}{2\eta_0}\hat{z} = \frac{\eta_0}{2}H_0^2\hat{z}, \tag{26.6}$$

where $\hat{z}$ is the direction of propagation of the wave. In later chapters, the energy carried by a monochromatic wave will be the starting point to understand the interaction of light with matter. In the next chapter, we will discuss how the energy carried by an electromagnetic wave as described by Equation 26.6 actually appears not in continuous quantities, but in quantum packets. Before we do that, we briefly discuss the classical picture of light interacting with material media.

26.4 Maxwell equations in materials

How does light interact with a material medium? Running the video of the process of the *creation* of light in Fig. 26.3 backwards, we can say that when an electromagnetic wave hits a metal wire, the electric field will slosh electrons in the wire back and forth, generating an AC current. This is the principle of operation of a receiving antenna. What happens when the material does not have freely conducting electrons like a metal? For example, in a dielectric, some electrons are tightly bound to atomic nuclei (core electrons), and others participate in forming chemical bonds with nearest neighbor atoms. The electric field of the electromagnetic wave will deform the electron clouds that are most "flexible" and "polarize" them. Before the external field was applied, the centroid of the negative charge from the electron clouds and the positive nuclei exactly coincide in space. When the electron cloud is deformed, the centroids do not coincide any more, and a net charge dipole is formed, as shown in Fig. 26.5. The electric field of light primarily interacts with electrons that are most loosely bound and deformable; protons in the nucleus are far heavier, and held strongly in place in a solid medium. Let's give these qualitative observations a quantitative basis.

The displacement vector in free space is $\mathbf{D} = \epsilon_0 \mathbf{E}$. In the presence of a dielectric, it has an additional contribution $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$, where $\mathbf{P}$ is the polarization of the dielectric. The classical model of polarization is an electric dipole $\mathbf{p}_i = q d_i \hat{n}$ in every unit cell of the solid where $\hat{n}$ is the unit vector along with the dipole points, and d_i is the charge separation. This dipole has zero magnitude in the absence of the external field[8]. The electric field of light stretches the electron cloud along it, forming dipoles along the direction of the field. Thus, $\mathbf{p}_i$ points along $\mathbf{E}$. The net polarization[9] is the volume-averaged dipole density $\mathbf{P} = \frac{1}{V}\sum_V \mathbf{p}_i$. Based on the material properties of the dielectric, we absorb all microscopic details into one parameter by writing

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}, \tag{26.7}$$

where the parameter χ_e is referred to as the electric susceptibility of the solid. With this definition, the displacement vector becomes

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \epsilon_0 \chi_e \mathbf{E} = \epsilon_0 (\underbrace{1 + \chi_e}_{\epsilon_r}) \mathbf{E} = \epsilon \mathbf{E}, \tag{26.8}$$

where the dielectric property of the material is captured by the modified dielectric constant $\epsilon = \epsilon_0 \epsilon_r = \epsilon_0(1 + \chi_e)$. The *relative* dielectric constant is 1 plus the electric susceptibility of the material. Clearly the relative dielectric constant of vacuum is 1, since there are no atoms to polarize and nothing is ”susceptible”.

In exactly the same way, if the material is magnetically polarizable, then $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$, where $\mathbf{M}$ is the magnetization vector. If there are tiny magnetic dipoles $\mathbf{m}_i = IA\hat{n}$ formed by circular loops carrying current I in area A in the material medium (see Fig. 26.5), the macroscopic magnetization is given by $\mathbf{M} = \frac{1}{V}\sum_V \mathbf{m}_i = \chi_m \mathbf{H}$, which leads to the relation

$$\mathbf{B} = \mu_0(\mathbf{H} + \chi_m \mathbf{H}) = \mu_0 (\underbrace{1 + \chi_m}_{\mu_r}) \mathbf{H} = \mu \mathbf{H}. \tag{26.9}$$

With these changes, the original Maxwell equations remain the same, but now $\mathbf{D} = \epsilon \mathbf{E}$ and $\mathbf{B} = \mu \mathbf{H}$, so we make the corresponding changes $\epsilon_0 \to \epsilon = \epsilon_0 \epsilon_r$ and $\mu_0 \to \mu = \mu_0 \mu_r$ everywhere. For example, the speed of light in a material medium then becomes $v = \frac{1}{\sqrt{\mu\epsilon}} = \frac{c}{\sqrt{\epsilon_r \mu_r}}$. If the material is non-magnetic, then $\mu_r = 1$, and $v = \frac{c}{\sqrt{\epsilon_r}} = \frac{c}{n}$, where $n = \sqrt{\epsilon_r}$ is called the refractive index of the material. Thus light travels slower in a material medium than in free space. Similarly, the wave impedance becomes $\eta_0 \to \eta = \sqrt{\frac{\mu}{\epsilon}} = \frac{\eta_0}{n}$ where the right equality holds for a non-magnetic medium[10].

If the material medium is conductive, or can absorb the light through electronic transitions, then the phenomena of *absorption* and corresponding attenuation of the light is captured by introducing an imaginary component to the dielectric constant, $\epsilon \to \epsilon_R + i\epsilon_I$. This leads to an imaginary component of the propagation vector **k**, which leads to attenuation. In Chapter 27 we calculate the absorption coefficients from quantum mechanics.

---

[8] Except in materials that have spontaneous, piezoelectric, or ferroelectric polarization.

[9] The classical picture of polarization as a dipole in every unit cell is not consistent with quantum mechanical Bloch states for electronic in crystals. The quantum theory of polarization requires the concept of the Berry phase which was discussed in Chapter 9.

[10] Light propagating in magnetic materials for which $\mu_r \neq 1$ exhibits fascinating properties such as Faraday rotation that was discussed in Chapter 25.

Electric and magnetic field lines may cross interfaces of different material media. Then, the Maxwell equations provide rules for tracking the magnitudes of the tangential and perpendicular components across interfaces. These **boundary conditions** are given by

$$\boxed{\begin{aligned} \mathbf{E}_{1t} - \mathbf{E}_{2t} &= 0, \\ \mathbf{H}_{1t} - \mathbf{H}_{2t} &= \mathbf{J}_s \times \hat{\mathbf{n}}, \\ D_{1n} - D_{2n} &= \rho_s, \\ B_{1n} - B_{2n} &= 0. \end{aligned}} \tag{26.10}$$

In words, the boundary condition relations say that the tangential component of the electric field $\mathbf{E}_t$ is *always* continuous across an interface, but the normal component is discontinuous if there are charges at the interface. If there are no *free* charges at the interface ($\rho_s = 0$), $\epsilon_1 E_{1n} = \epsilon_2 E_{2n}$, implying the normal component of the electric field is larger in the material with a smaller dielectric constant. This feature is used in Si MOSFETs, where much of the electric field drops across an oxide layer rather than in the semiconductor which has a higher dielectric constant. Similarly, the normal component of the magnetic field is always continuous across an interface, whereas the tangential component can change if there is a surface current flowing at the interface of the two media.

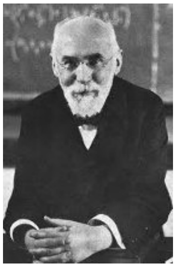


Fig. 26.6 Hendrik Lorentz (physics Nobel prize in 1902) was the original architect of the "Lorentz transformations" that led to Einstein's theory of relativity. He contributed to all areas of physics, including the Drude-Lorentz classical model for light-matter interaction.

The force (in newtons) on a particle of charge q in the presence of an electric and magnetic field is given by the Lorentz (Fig. 26.6) equation

$$\boxed{\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}).} \tag{26.11}$$

Since the energy of the charged particle changes as $W = \int \mathbf{F} \cdot d\mathbf{r}$, the rate of change of energy is $\mathbf{F} \cdot \mathbf{v} = q\mathbf{E} \cdot \mathbf{v}$, which is the power delivered to the charged particle by the fields. Note that a static magnetic field cannot deliver power since $\mathbf{v} \times \mathbf{B} \cdot \mathbf{v} = 0$. Thus a time-independent magnetic field cannot change the energy of a charged particle. But a time-dependent magnetic field creates an electric field, which can.

Table 26.1 Relation between real and imaginary parts of dielectric constants and refractive indices and the optical reflectivity.

$\eta + i\kappa = \sqrt{\epsilon_1 + i\epsilon_2}$
$\epsilon_1 = \eta^2 - \kappa^2$ and $\epsilon_2 = 2\eta\kappa$ $\eta = \sqrt{\frac{\sqrt{\epsilon_1^2+\epsilon_2^2}+\epsilon_1}{2}}$ and $\kappa = \sqrt{\frac{\sqrt{\epsilon_1^2+\epsilon_2^2}-\epsilon_1}{2}}$
$R = \frac{(\eta-1)^2+\kappa^2}{(\eta+1)^2+\kappa^2}$

26.5 Classical light-matter interaction

In Chapter 2, Exercise 2.5 we encountered the Drude model for interaction of light with matter to explain why metals are shiny. Then, we discussed a similar model to explain the Reststrahlen band that results from the interaction of light with the optical phonon modes of semiconductors in Chapter 22, Section 22.3. In particular, by picturing phonons as oscillators interacting with the electromagnetic light wave, the dielectric constant was derived. That model in its most generalized form accounts for resonances not just from phonons, but free carriers, and also interband transitions. This generalized **Drude–Lorentz** model for the frequency-dependent relative dielectric constant is

$$\boxed{\epsilon_r(\omega) = 1 + \Sigma_j \frac{\omega_{p,j}^2}{\omega_j^2 - \omega^2 - i\gamma_k \omega}.} \tag{26.12}$$

In this expression each resonance is modeled as the electric field of light interacting with N oscillators per unit volume. Each oscillator type has its respective effective charge and mass, forming an effective plasmon resonance frequency $\omega_{p,j}$ given by $\omega_{p,j}^2 = f_j(N(q_j^\star)^2)/(\epsilon_0 m_j^\star)$. The natural resonance frequency of the j^{th} resonance is ω_j, and the oscillator strength for that resonance is f_j, which satisfies the sum rule $\sum_j f_j = 1$. Each resonance is characterized by a separate damping γ_j. The resulting real and imaginary part of the refractive index $\eta + i\kappa$ can be evaluated from the **Sellmeier equation** using the standard relations reproduced in Table 26.1. Fig. 26.7 shows the resulting refractive indices. In addition to the phonon resonance in the IR, the free carrier resonance and interband resonances also appear at higher frequencies of light. The imaginary part of the resonances result in absorption of light in the medium, with the intensity reducing from the surface of the solid as $I(z) \sim I_0 e^{-2\kappa z}$, indicating that $\kappa(\omega)$ is related to the **absorption coefficient** of the solid. In Chapter 27, we will discuss the absorption coefficient of semiconductors from the quantum mechanical picture, which will add several features to the classical model. It is worthwhile comparing Fig. 26.7 (b) to Fig. 27.4 for a preview of this connection.

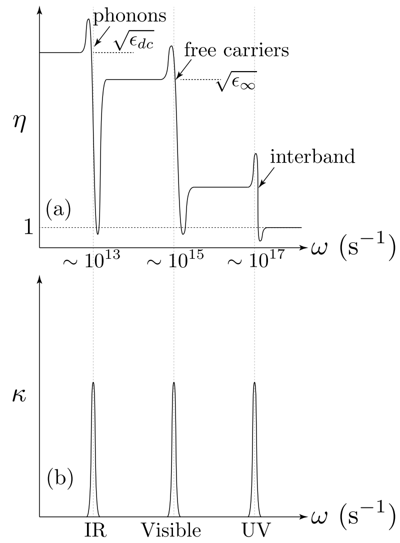


Fig. 26.7 The real part η and imaginary part κ of the refractive index of solids resulting from resonances in light-matter interaction. The regions of high κ define the absorption bands.

26.6 Kramers–Kronig relations

The relation between the frequency dependent refractive index and dielectric constant of a material is $\eta(\omega) + i\kappa(\omega) = \sqrt{\epsilon_1(\omega) + i\epsilon_2(\omega)}$. But the real and imaginary parts are themselves related. Using the mathematical properties of complex functions and the physical constraint of causality, the frequency dependent real and imaginary parts of the refractive index are related by the **Kramers–Kronig** relations for the dielectric function[11]:

$$\epsilon_1(\omega) = 1 + \frac{2}{\pi}\mathcal{P}\int_0^\infty d\omega' \frac{\omega' \epsilon_2(\omega')}{\omega'^2 - \omega^2}$$
$$\epsilon_2(\omega) = -\frac{2\omega}{\pi}\mathcal{P}\int_0^\infty d\omega' \frac{\epsilon_1(\omega')}{\omega'^2 - \omega^2}. \tag{26.13}$$

The relations imply that if the real part $\epsilon_1(\omega)$ or the imaginary part $\epsilon_2(\omega)$ is measured, the other may be directly extracted using the KK relation and does not require another measurement. This would be of immense practical use, as will be discussed in Chapter 27, where the real part of the dielectric function may be completely concluded by measuring $\epsilon_2(\omega)$ via absorption spectroscopy. Similarly, for the

[11] Here $\mathcal{P}$ means the principal part of the integral, which is evaluated by adding a small imaginary component in the denominator and taking its zero limit.

frequency dependent refractive index the Kramers–Kronig relations are

$$\eta(\omega) = 1 + \frac{1}{\pi}\mathcal{P}\int_{-\infty}^{+\infty} d\omega' \frac{\kappa(\omega')}{\omega' - \omega}$$
$$\kappa(\omega) = -\frac{1}{\pi}\mathcal{P}\int_{-\infty}^{+\infty} d\omega' \frac{\eta(\omega') - 1}{\omega' - \omega}. \tag{26.14}$$

In principle, the use of the Kramers-Kronig relation requires the knowledge of either the real or imaginary parts over the *entire* frequency range from $0 \leq \omega \leq \infty$. In practice, a fit to experimentally measured data over a sufficient frequency range is used.

26.7 Accelerating charges radiate

When a point charge is *accelerated* with acceleration a, it radiates electromagnetic waves. Radiation travels at the speed of light. So the electric and magnetic fields at a point far from the charge are determined by a *retarded* response. Using retarded potentials, or more intuitive approaches[12], one obtains that the radiated electric field goes as

$$\mathbf{E}_r = (\frac{qa}{4\pi\epsilon_0 c^2})\frac{\sin\theta}{r}\hat{\theta}, \tag{26.15}$$

expressed in spherical coordinates with the charge at the origin, and accelerating along the x-axis. Note that here the subscript r stands for radiated, it is not the radial component. The radiated magnetic field $\mathbf{H}_r$ curls in the $\hat{\phi}$ direction and has a magnitude $|\mathbf{E}_r|/\eta_0$. The radiated power is obtained by the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ as

$$\mathbf{S} = (\frac{\mu_0 q^2 a^2}{16\pi^2 c^2})(\frac{\sin\theta}{r})^2\hat{r}. \tag{26.16}$$

Note that unlike static charges or currents that fall as $1/r^2$ away from the source, the *radiated* **E** and **H** fields fall as $1/r$. If they didn't, the net power radiated very far from the source will go to zero since $\oint \mathbf{S}\cdot d\mathbf{A} \sim S(r)4\pi r^2 \to 0$. Integrating the power over the angular coordinates results in the famous **Larmor Formula** for the net electromagnetic power in Watts radiated by an accelerating charge:

$$\boxed{P = \frac{\mu_0 q^2 a^2}{6\pi c} = \frac{q^2 a^2}{6\pi\epsilon_0 c^3}.} \tag{26.17}$$

26.8 Need for quantum theory of light

Classical electromagnetism contained in Maxwell's equations can explain a remarkably large number of experimentally observed phenomena, but not all. We discussed in the beginning of this chapter that radiation of electromagnetic waves can be created in an antenna, which

[12] An intuitive picture for radiation by an accelerating charge was first given by J. J. Thomson, the discoverer of the electron. A relativistic derivation gives the same final result as Equation 26.17, but requires the charge to undergo non-uniform acceleration – such as a charge undergoing periodic oscillations – as is the case in a radiating antenna.

in its most simple form is a conducting wire in which electrons are sloshed back and forth. The *collective* acceleration, coupled with the Larmor formula can explain radiation from a vast number of individual sources of electromagnetic radiation.

As we discussed in Chapter 3, by the turn of the 20th century, improvements in spectroscopic equipment had helped resolve what was originally thought as broadband (many frequencies ω) radiation into the purest spectral components. It was observed that different gases had different spectral signatures. The most famous among them were the spectral features of the hydrogen atom, then known as the hydrogen gas. There is nothing *collective* about hydrogen gas, since it is not a conductor and there are not much electrons to slosh around as a metal – there is only *one* electron! The classical theory for radiation proved difficult to apply to explain the spectral features. Classical electromagnetism could not explain the photoelectric effect, and the spectrum of blackbody radiation either. The search for an explanation led to the quantum theory of light, which is the subject of the next chapter.

26.9 Chapter summary section

In this chapter, we learned:

- Light is an electromagnetic wave with oscillating **E** and **H** fields propagating in a direction perpendicular to both fields at a speed $c \sim 3 \times 10^8$ m/s in vacuum. The energy per unit area per unit time carried by light is given by the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$.
- The propagation properties of light in vacuum as well as in material media is explained by the four Maxwell's equations.
- Light interacts with material media via its electric and magnetic fields through several resonances with the charges and spins in the electrons and nuclei. These resonances lead to the phenomena of reflection, transmission, absorption, and scattering.
- Maxwell equations assume the energy in light to be continuous. This model fails to explain blackbody radiation and emission spectra of atoms and molecules, for which light is modeled as a collection of quantum harmonic oscillators with discrete energies.

Further reading

There are a wealth of excellent books on electromagnetic theory. I will mention some that I have studied or referred to over the years. **Introduction to Electrodynamics** by D. Griffiths is an excellent introductory textbook to learn the subject. **Classical Electrodynamics** by J. D. Jackson completes the subject by elevating the mathematical structure and clarifying how close the classical theory comes to explaining the electromagnetic phenomena that need the machinery of quantum theory. **The Quantum Theory of Light** by R. Loudon highlights the beautiful structure and modern aspects of the quantum nature of light. **Optical Properties of Solids** by M. Fox is a very readable and comprehensive textbook on light-matter interactions, which describes a wealth of experimental methods. **Electromagnetic Fields and Waves** by M. Iskander and **Fields and Waves in Communication Electronics** by S. Ramo, J. Whinnery and T. Van Duzer treat the several engineering aspects of electromagnetic waves ranging from antennas and microwave propagation in transmission lines and circuits.

Exercises

(26.1) **Survey of electromagnetic waves and the impact of Maxwell equations**

As the only exercise of this introductory chapter of Module IV, write a short essay of the profound effect of our understanding of light and electromagnetic waves via the Maxwell equations.

Light–Matter Interaction

27

27.1 Photonic effects: a résumé 717
27.2 Electron-photon matrix elements 719
27.3 Absorption spectra of semiconductors 722
27.4 Number of photons in light 723
27.5 Photon absorption rate 724
27.6 Equilibrium absorption coefficient 729
27.7 Quantum wells, wires, and dots 731
27.8 Critical points 734
27.9 Forbidden and indirect absorption 738
27.10 Exciton absorption 740
27.11 Franz–Keldysh effect 742
27.12 Intersubband absorption 744
27.13 Free carrier, and impurity absorption 745
27.14 Photoelectron spectroscopy 747
27.15 Chapter summary section 747
Further reading 748
Exercises 748

In this chapter, we explore the fundamental optical transitions that occur in semiconductors. The reader is strongly encouraged to first revise the summary of photonic properties of semiconductors in Chapter 16 (from Section 16.9 till the end of that chapter) to have the "forest" in full view as we dive deeper into the trees (or the details) in this and following chapters. The specific questions addressed in this chapter are:

- How can we capture the interaction of semiconductors with photons quantum mechanically?
- How can we explain the optical absorption spectra of bulk semiconductors in equilibrium?
- How are these properties modified in low-dimensional semiconductor structures such as quantum wells and quantum dots?
- What are the characteristics of optical processes other than interband transitions such as absorption by free-carriers, excitons, impurities, and intersubband transitions, and how do they differ from interband transitions?

27.1 Photonic effects: a résumé

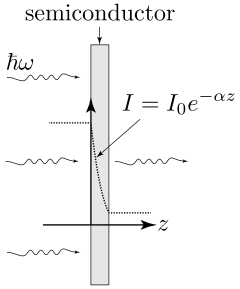


Fig. 27.1 By varying the incoming photon energy $\hbar\omega = h\nu$, the absorption coefficient $\alpha(\hbar\omega)$ of a semiconductor is measured. The semiconductor also reflects and scatters light, and can produce photoluminescence.

Fig. 27.1 shows a popular experiment used to probe the strength of the interaction of light with any material. Historically absorption spectroscopy measurements of this kind directly provide the absorption coefficient $\alpha(\hbar\omega)$ as a function of the photon energy $\hbar\omega$. They were the earliest probes of the energy bandgap and the internal electronic bandstructure of semiconductors, and remain the most popular today. In addition to the **absorption**, and **transmission** of the part not absorbed, light is also **reflected**, **scattered**, or re-emitted as **luminescence**. Each of these optical processes provide information about the electronic and vibrational properties of the semiconductor. Using ultrashort pulses of light, time-resolved spectroscopy of each kind enables the measurement of the rates of interband, intersubband, or defect-induced transitions. It is useful to draw the analogy of light-matter interaction with how electrons interact with phonons, which was discussed in Chapter 22. The mathematical treatment of light-matter interaction is nearly identical to electron-phonon interactions, but there are very important differences as well.

Electrons in a semiconductor can absorb and emit photons. The

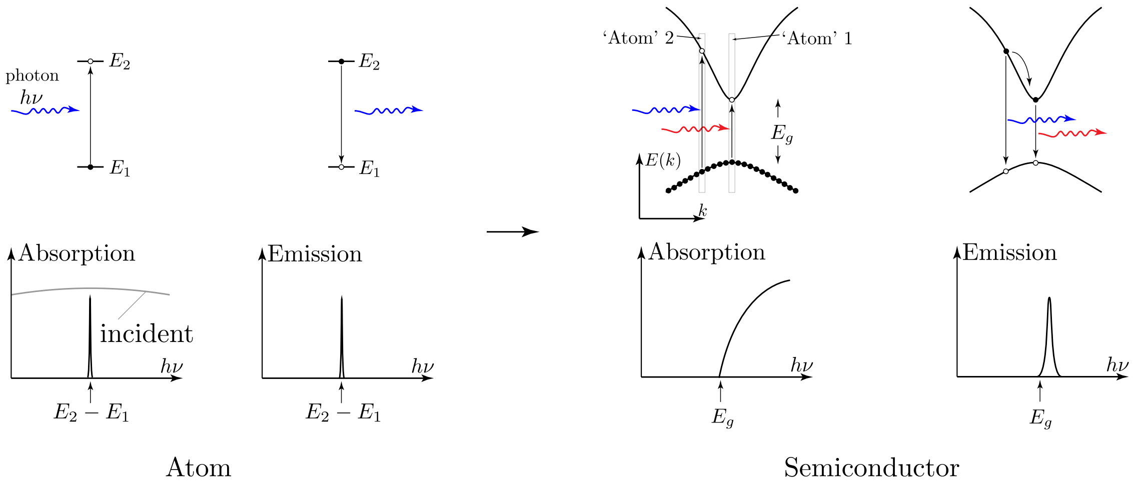


Fig. 27.2 Analogy of optical properties of atoms and of semiconductors.

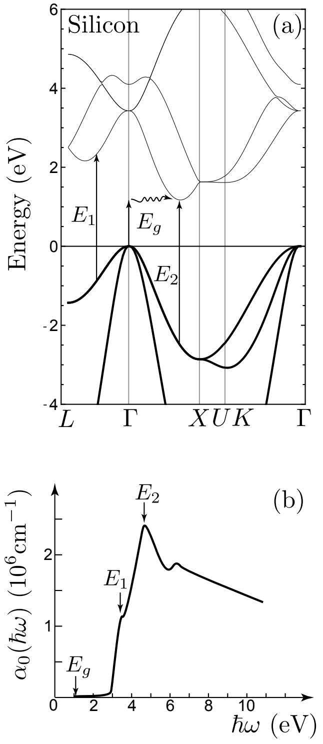


Fig. 27.3 (a) Silicon electron bandstructure calculated by the tight-binding method in Chapter 11, and (b) its interband optical absorption spectrum.

emission can be spontaneous or stimulated. In Chapter 16, Section 16.9 we got a sneak peek of the optical transitions in semiconductors and the comparison to their counterpart in atoms. Fig. 27.2 shows the processes (again). Because of the sharp electron energy eigenvalues E_1, E_2 of an atom, the absorption and emission lines at $\hbar\omega = h\nu = E_2 - E_1$ are very sharp, and have a characteristic lineshape that depends on the rate of the transition. As we will discuss in this chapter, because of the presence of energy bands in the semiconductor, the absorption spectrum depends on the **optical joint density of states**. The lineshape of the equilibrium absorption spectrum due to interband transitions $\alpha_0(\hbar\omega)$ is characteristic of the particular light-matter interaction process. For near band-edge absorption it varies as $\alpha_0(\hbar\omega) \propto (\hbar\omega - E_g)^n$ where E_g is the bandgap of the semiconductor and n is characteristic of the process: allowed vs. forbidden, or direct vs. indirect, and has a strong fingerprint of the dimensionality of carriers. Fig. 27.3 shows the absorption spectrum of silicon with its bandstructure. The indirect transition of silicon at the bandgap $\hbar\omega = E_g$ requires a phonon and is weak and has a low absorption coefficient. But the direct transitions E_1 and E_2 indicated are strong and result in a high absorption coefficient.

The absorption spectrum changes with temperature or doping of the semiconductor. It changes dramatically when the semiconductor is pulled out of equilibrium. Interband absorption processes need not be direct: indirect absorption is weaker because it requires the presence of phonons, but is of immense practical value (e.g. in silicon solar cells). We will see that interband transitions are not the only processes by which semiconductors absorb photons. Processes of **free-carrier**, **defect-assisted**, and **excitonic** transitions are important. In semiconductor **quantum wells** and **quantum wires**, intersubband absorption is also possible, and is of practical value.

Direct-bandgap semiconductors produce strong luminescence by interband transitions as indicated in Fig. 27.2. Indirect bandgaps produce weak interband luminescence. Both direct and indirect bandgap semiconductors can produce luminescence due to transitions of electrons in the conduction or valence bands to defect states located inside the bandgap. Semiconductors do not need to be excited by light to produce luminescence: they can directly convert electrical current into light. Indeed, this is the basis for **light-emitting diodes** and **lasers**. Lasers can be based on both interband and intersubband transitions. Since semiconductor bandgaps range from near $0 - 6$ eV, the spectral range covered in semiconductor photonic devices ranges from the long-wavelength **IR to deep-ultraviolet**. As a rigorous preparation for discussing photonic devices such as solar cells and photodetectors in Chapter 28 and LEDs and lasers in Chapter 29, in this chapter we explore the quantum physics of these light-matter interaction processes.

27.2 Electron-photon matrix elements

We now develop the quantum mechanics of optical transitions in semiconductors[1]. The energy of the electron in a semiconductor is modified by the presence of light. For an electromagnetic wave represented by its magnetic vector potential $\mathbf{A}$ and electron charge $-q$, the electron-photon interaction is captured in the Hamiltonian

$$\boxed{\hat{H} = \frac{1}{2m_e}(\hat{\mathbf{p}} + q\mathbf{A})^2 + V(\mathbf{r}),} \tag{27.1}$$

where $\hat{\mathbf{p}}$ is the quantum mechanical momentum operator, m_e is the electron rest-mass (*not* effective mass!), and $V(\mathbf{r})$ is the potential experienced by the electron. This form of the light-matter Hamiltonian is semi-classical in the sense that the electromagnetic wave is represented by $\mathbf{A}$, the classical vector potential of Maxwell's equation. Though it is possible to use the quantized version of the electromagnetic field[2], we delay doing so till Chapter 29. Every physical phenomenon we will discuss (except spontaneous emission) is explained completely by the vector potential approach. The mathematical treatment is a straightforward application of perturbation theory and Fermi's golden rule.

The above Hamiltonian completely solves the problem of the interaction of light with atoms and explains, say, the sharp absorption and emission spectra of a hydrogen atom[3]. Here, we apply this light-matter Hamiltonian to semiconductors, in which the energy eigenvalues form bands separated by energy gaps.

The vector potential $\mathbf{A}$ is related to the electric field (see Chapter 26):

$$\nabla \times \mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) \underbrace{\rightarrow}_{\mathbf{B}(\mathbf{r},t)=\nabla\times\mathbf{A}(\mathbf{r},t)} \mathbf{E}(\mathbf{r},t) = -\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r},t), \tag{27.2}$$

and we choose to work[4] in the Coulomb gauge in which $\nabla \cdot \mathbf{A} = 0$.

[1] The reader is encouraged to review Section 22.8 to draw analogies between the physics of electron-photon interactions with electron-phonon interactions.

[2] The quantized version of the electromagnetic field is necessary at the extreme low intensity limit of single photons. It is discussed in Chapter 29.

[3] For the hydrogen atom $V(r)$ of Equation 27.1 will be the Coulomb potential of the proton.

[4] The popular statement: *"I have never read a paper that states that this experiment was performed in the Coulomb gauge"* emphasizes that the physically measurable quantities do not depend on the choice of gauge.

This strategic choice allows the vector potential **A** to commute with the momentum operator

$$[\hat{\mathbf{p}}, \mathbf{A}] = 0 \to \hat{\mathbf{p}} \cdot \mathbf{A} = \mathbf{A} \cdot \hat{\mathbf{p}}, \tag{27.3}$$

which leads to the electron-photon Hamiltonian

$$\hat{H} = \underbrace{[\frac{\hat{p}^2}{2m_e} + V(\mathbf{r})]}_{\hat{H}_{el}} + \underbrace{\frac{q}{m_e}\mathbf{A} \cdot \hat{\mathbf{p}}}_{\hat{W}_{el-photon}} + \underbrace{\frac{q^2 A^2}{2m_e}}_{\text{non-linear effects}} . \tag{27.4}$$

The light-matter Hamiltonian is seen as a sum of the electron Hamiltonian $\hat{H}_{el}$, and two "perturbation" terms[5] seen by the electron due to the electromagnetic wave (the second and third terms vanish for $A = 0$). For an electron in a semiconductor crystal, the potential energy term in the unperturbed Hamiltonian is the periodic crystal potential $V(\mathbf{r} + \mathbf{a}_0) = V(\mathbf{r})$, where $\mathbf{a}_0$ is a lattice constant. We neglect the perturbation term that goes as the square of the magnetic vector potential for "weak" intensities of light. This is justified when the condition $|q\mathbf{A}| << |\mathbf{p}| \sim \hbar\pi/a_0$ is met; in other words, we neglect the term $(qA)^2/(2m_e)$ with respect to $\hat{p}^2/(2m_e)$. The net Hamiltonian we retain then has the electron experiencing the perturbation

$$\boxed{\hat{W}_{el-photon} = \frac{q}{m_e}\mathbf{A} \cdot \hat{\mathbf{p}},} \tag{27.5}$$

due to its interaction with light. The magnetic vector potential **A** for an electromagnetic (= EM) wave is of the form

$$\mathbf{A}(\mathbf{r}, \mathbf{t}) = \hat{e}A_0 \cos(\mathbf{k_{op}} \cdot \mathbf{r} - \omega t) = \hat{e}\frac{A_0}{2}e^{+i\mathbf{k_{op}}\cdot\mathbf{r}}e^{-i\omega t} + \hat{e}\frac{A_0}{2}e^{-i\mathbf{k_{op}}\cdot\mathbf{r}}e^{+i\omega t} \tag{27.6}$$

where ω is the angular frequency of the EM wave, $\hat{e}$ is the unit vector along the electric (and vector potential) field, and $\mathbf{k_{op}}$ is the propagation wave vector of magnitude $2\pi/\lambda$. The electron-photon interaction Hamiltonian is then given by

$$\hat{W}_{el-photon}(\mathbf{r}, t) = \frac{q}{m_e}\mathbf{A} \cdot \hat{\mathbf{p}} = \hat{W}_{+}(\mathbf{r})e^{-i\omega t} + \hat{W}_{-}(\mathbf{r})e^{+i\omega t}$$

$$\implies \boxed{\hat{W}_{+}(\mathbf{r}) = \frac{qA_0 e^{i\mathbf{k_{op}}\cdot\mathbf{r}}}{2m_e}\hat{e} \cdot \hat{\mathbf{p}}} \text{ and } \boxed{\hat{W}_{-}(\mathbf{r}) = \frac{qA_0 e^{-i\mathbf{k_{op}}\cdot\mathbf{r}}}{2m_e}\hat{e} \cdot \hat{\mathbf{p}},} \tag{27.7}$$

where $\hat{W}_{+}$ stands for photon absorption and $\hat{W}_{-}$ for photon emission, as may be concluded by direct analogy to the phonon case in Section 22.8. This perturbation term is now used in Fermi's golden rule to evaluate the transition rates for electrons due to the interaction with photons.

An alternative formulation to the $\hat{W}_{el-photon} = \frac{q}{m_e}\mathbf{A} \cdot \hat{\mathbf{p}}$ perturbation is popular for the light-matter interaction Hamiltonian. It is given by

---

[5] The solution of the first term $\hat{H}_{el}$ is, of course, the entire bandstructure $E_n(\mathbf{k})$ of the semiconductor. Note that if the second and third terms can be treated as perturbations for low light intensity. The discussion in this chapter is limited to such perturbative effects. For strong coupling, quasiparticles are formed, characterized by the eigenstates of the entire Hamiltonian. It is also useful to note that $|q\mathbf{A}|$ has units of momentum.

$\hat{W}_{el-photon} = q\mathbf{E}\cdot\mathbf{r}$, where $\mathbf{E}$ is the electric field of the EM wave, and $\mathbf{r}$ is the real-space coordinate: it is *not* an operator for the wavefunctions in real-space, it simply multiplies. We have encountered a variant of this in Chapter 9, Equation 9.38; let's see how this comes about again. The momentum operator in terms of the total Hamiltonian is

$$p_x = m_e \frac{\partial}{\partial t}\langle x\rangle = \frac{-im_e}{\hbar}[x, H]. \tag{27.8}$$

Since $E_b - E_a = \hbar\omega_{ba}$ and $\mathbf{E} = i\omega\mathbf{A}$ for a monochromatic EM wave, the momentum matrix element between states $|a\rangle$ and $|b\rangle$ is

$$\begin{gathered}(p_{ba})_x = \langle b|p_x|a\rangle = \frac{-im_e}{\hbar}\langle b|xH - Hx|a\rangle = \frac{im_e}{\hbar}(E_a - E_b)\langle b|x|a\rangle \\ (p_{ba})_x\hat{x} + (p_{ba})_y\hat{y} + (p_{ba})_z\hat{z} = \frac{im_e}{\hbar}(E_a - E_b)\langle b|x\hat{x} + y\hat{y} + z\hat{z}|a\rangle \\ \implies \langle b|\hat{\mathbf{p}}|a\rangle = \mathbf{p}_{ba} = \frac{im_e}{\hbar}(E_a - E_b)\mathbf{r}_{ba} \implies \boxed{\mathbf{p}_{ba} = im_e\omega_{ba}\mathbf{r}_{ba}.}\end{gathered} \tag{27.9}$$

This relation connects the momentum matrix element $\mathbf{p}_{ba}$ to the "dipole length" matrix element $\mathbf{r}_{ba}$ in a nice way for the resonant case $\omega = \omega_{ba}$. The electron-photon matrix element then becomes

$$\begin{gathered}W_{ab} = \langle b|\frac{q}{m_e}\mathbf{A}\cdot\hat{\mathbf{p}}|a\rangle = \frac{q}{m_e}\mathbf{A}\cdot\langle b|\hat{\mathbf{p}}|a\rangle = \frac{q}{m_e}\mathbf{A}\cdot im_e\omega_{ba}\langle b|\mathbf{r}|a\rangle = q\mathbf{E}\cdot\langle b|\mathbf{r}|a\rangle \\ \implies \boxed{\frac{q}{m_e}\mathbf{A}\cdot\hat{\mathbf{p}} = q\mathbf{E}\cdot\mathbf{r}} \text{ inside the optical matrix element.}\end{gathered} \tag{27.10}$$

The $\mathbf{E}\cdot\mathbf{r}$ form can be used interchangeably with the $\mathbf{A}\cdot\hat{\mathbf{p}}$ form depending on the problem for resonant excitation $\omega = \omega_{ba}$.

The electron-photon matrix elements to be used in Fermi's golden rule for semiconductors are of the form $\langle\mathbf{k_c}|\hat{W}_+|\mathbf{k_v}\rangle$ and $\langle\mathbf{k_c}|\hat{W}_-|\mathbf{k_v}\rangle$, where the *unperturbed* electron states $|\mathbf{k_c}\rangle$ and $|\mathbf{k_v}\rangle$ are solutions of the unperturbed Hamiltonian $H_{el} = \frac{\hat{p}^2}{2m_e} + V(\mathbf{r})$. But this Hamiltonian yields precisely the electron bandstructure of the semiconductor. The electronic states in the effective mass approximation are $\psi_c(\mathbf{r}) = \langle\mathbf{r}|\mathbf{k_c}\rangle = C_c(\mathbf{r})u_c(\mathbf{r})$ for the conduction band and $\psi_v(\mathbf{r}) = \langle\mathbf{r}|\mathbf{k_v}\rangle = C_v(\mathbf{r})u_v(\mathbf{r})$ for the valence band. Here $C_c(\mathbf{r})$ and $C_v(\mathbf{r})$ are the slowly varying envelope functions, and $u_c(\mathbf{r})$ and $u_v(\mathbf{r})$ are the the Bloch lattice functions. Near the band-edge of direct-gap semiconductors, $u_c(\mathbf{r})$ is primarily s-orbitals and $u_v(\mathbf{r})$ is primarily p-orbitals. With the proper choice of the envelope functions, this model applies to quantum wells, wires, or dots, which will be discussed later. For bulk 3D semiconductors, the electronic states take the form $\psi_c(\mathbf{r}) = (\frac{e^{i\mathbf{k_c}\cdot\mathbf{r}}}{\sqrt{V}})u_c(\mathbf{r})$ and $\psi_v(\mathbf{r}) = (\frac{e^{i\mathbf{k_v}\cdot\mathbf{r}}}{\sqrt{V}})u_v(\mathbf{r})$ where $\mathbf{k}_c$ and $\mathbf{k}_v$ are the electron and hole wavevectors, and V is the crystal volume. The effective mass approximation transforms the unperturbed electronic Hamiltonian into the

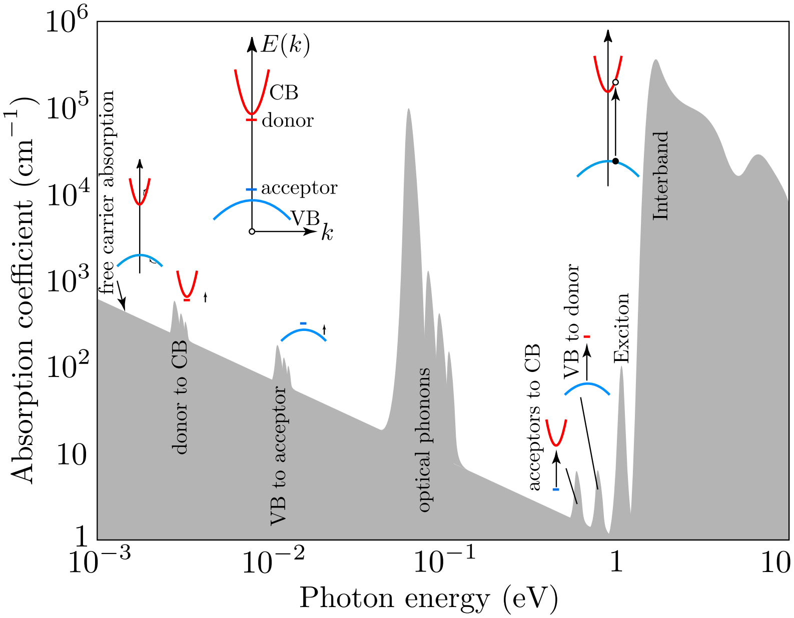


Fig. 27.4 Absorption spectrum of semiconductors. A schematic absorption spectrum $\alpha(\hbar\omega)$ of bulk semiconductors. The insets depict various state transitions upon absorption of photons. Adapted from *Physical Properties of Semiconductors* by Wolfe, Holonyak, and Stillman.

much simpler form (e.g. $\hat{p}^2/(2m_e) + V(\mathbf{r}) \to \hat{p}^2/2m_c^\star$ for the electron), and the corresponding effective-mass Schrödinger equation is $\hat{p}^2/(2m^\star)\psi_c(\mathbf{r}) = (E - E_c)\psi_c(\mathbf{r})$ for the conduction band. We will work in this effective-mass theory. The advantage of working in the effective-mass theory is that the light-matter interaction matrix elements for electrons confined in low-dimensional structures such as quantum wells, wires, or dots follows in a simple way from the bulk results. We will need the matrix elements shortly to explain the absorption spectra of bulk semiconductors, which we discuss next.

27.3 Absorption spectra of semiconductors

The **Beer–Lambert** law states that if light of intensity I_0 is incident on a material that absorbs, the intensity will decay inside the material as $I(z) = I_0 e^{-\alpha z}$. Here α is the absorption coefficient, in units of inverse length. Typically the unit used for α is cm^{-1}. Let us consider the following experiment: take a piece of bulk semiconductor, say GaAs or GaN, and using a tunable light source, measure α as a function of the photon energy $\hbar\omega$. Then we obtain the absorption spectrum $\alpha(\hbar\omega)$. The absorption spectrum of most semiconductors looks like what is shown in the schematic Fig. 27.4.

The insets of Fig. 27.4 indicates the electron bandstructure of the bulk semiconductor, including states corresponding to donor and acceptor dopants. The transitions between electron states caused by photon absorption are indicated. The decreasing floor of the absorption spectrum up to $\sim$ 1 eV is due to intraband transitions caused by the absorption of low energy (∼ few meV) photons by free carriers. Transitions between dopant and band states are shown, in addition to the below-bandgap excitonic transition. Such optical measurements provide a sensitive experimental determination of dopant and excitonic energies with respect to the fundamental band edge energies. Photons can excite phonons in the bulk semiconductor crystal: the absorption peak for this process is rather strong, and forms the basis of Raman spectroscopy, as was discussed in Chapter 22.

By far, the strongest absorption occurs for interband transitions, which is the focus of the next few sections. The absorption spectrum is quantitatively defined as

$$\boxed{\alpha(\hbar\omega) = \frac{\text{\# of photons absorbed per unit volume per second}}{\text{\# of photons incident per unit area per second}} = \frac{R_{abs}(\hbar\omega)}{N_{ph}(\hbar\omega)}.} \tag{27.11}$$

We derive an expression for the denominator $N_{ph}(\hbar\omega)$ in the next section, and for the numerator $R_{abs}(\hbar\omega)$ in the section thereafter.

27.4 Number of photons in light

Consider a monochromatic EM wave of frequency ω and corresponding wavevector $\mathbf{k_{op}} = \frac{2\pi}{\lambda}\hat{n}$. For a plane wave, the magnetic vector potential is

$$\mathbf{A}(\mathbf{r},\mathbf{t}) = \hat{e}A_0\cos(\mathbf{k_{op}}\cdot\mathbf{r} - \omega t), \tag{27.12}$$

from where the electric field is obtained by using

$$\begin{aligned}\mathbf{E}(\mathbf{r},t) &= -\frac{\partial}{\partial t}\mathbf{A}(\mathbf{r},t)\\ &= -\hat{e}\omega A_0\sin(\mathbf{k_{op}}\cdot\mathbf{r} - \omega t),\end{aligned} \tag{27.13}$$

and the magnetic field intensity is

$$\begin{aligned}\mathbf{H}(\mathbf{r},t) &= \frac{1}{\mu}\nabla\times\mathbf{A}(\mathbf{r},t)\\ &= -\frac{1}{\mu}\mathbf{k_{op}}\times\hat{e}A_0\sin(\mathbf{k_{op}}\cdot\mathbf{r} - \omega t).\end{aligned} \tag{27.14}$$

Here we have used $\nabla\times(...) \equiv -\mathbf{k_{op}}\times(...)$ for plane monochromatic waves, as described in Chapter 26. Then, the energy carried by the plane wave per unit area per unit time is given by the Poynting vector

$$\begin{aligned}\mathbf{S}(\mathbf{r},t) &= \mathbf{E}(\mathbf{r},t)\times\mathbf{H}(\mathbf{r},t)\\ &= \mathbf{k_{op}}\frac{\omega A_0^2}{\mu}\sin^2(\mathbf{k_{op}}\cdot\mathbf{r} - \omega t),\end{aligned} \tag{27.15}$$

where we use the identity $\hat{e}\times\mathbf{k_{op}}\times\hat{e} = \mathbf{k_{op}}$. Since the frequency of typical UV-visible-IR light is very high, we time-average the Poynting vector over a period to obtain

$$\langle \mathbf{S}(\mathbf{r},t)\rangle = \frac{\omega A_0^2}{2\mu}\mathbf{k_{op}}, \tag{27.16}$$

and its magnitude is

$$S = |\langle \mathbf{S}(\mathbf{r},t)\rangle| = \frac{\omega A_0^2}{2\mu}k_{op} = \frac{n_r c\epsilon_0 \omega^2 A_0^2}{2} = \frac{E_0^2}{2\eta}, \tag{27.17}$$

where $\mu = \mu_0$ and $n_r = \sqrt{\mu_r \epsilon_r}$ is the refractive index of the media, in which the speed of light is c/n_r. Also note[6] that $E_0 = \omega A_0$, and $\eta = \sqrt{\mu/\epsilon}$ is the field impedance. This relation gives us a way to find the magnitude of the vector potential A_0 if we know the power carried per unit area by the electromagnetic wave. Since energy in electromagnetic waves is carried in quantum packets (photons) of individual energy $\hbar\omega$, the **number of photons** that cross unit area per unit time is then given by

$$\boxed{N_{ph}(\hbar\omega) = \frac{S}{\hbar\omega} = \frac{n_r c\epsilon_0 \omega^2 A_0^2}{2\hbar\omega} = \frac{E_0^2}{2\eta\hbar\omega}.} \tag{27.18}$$

The **intensity** of light is proportional to the **square** of the electric (or magnetic) field amplitude, and thus the number of photons is a measure of the intensity of radiation. Equation 27.18 provides the denominator of the expression for absorption coefficient Equation 27.11. The numerator term is discussed in the next section.

27.5 Photon absorption rate

To find the rate of photon absorption in the bulk semiconductor, we apply Fermi's golden rule. We first note that the numerator of Equation 27.11 has units of number of photons absorbed per unit volume per second. Consider Fig. 27.5. An electron in the valence band state $|a\rangle$ absorbs a photon of energy $\hbar\omega$ and transitions into state $|b\rangle$ in the conduction band. Each such transition event results in the **annihilation of exactly one photon** from the EM field. The rate at which this happens is given by Fermi's golden rule as

$$\frac{1}{\tau_{a\to b}} = \frac{2\pi}{\hbar}|\langle b|W_{ba}(\mathbf{r})|a\rangle|^2 \delta[E_b - (E_a + \hbar\omega)], \tag{27.19}$$

where $\langle b|W(\mathbf{r})|a\rangle$ is the perturbation matrix element, and the Dirac delta function is a statement of energy conservation in the process. The reverse process of photon emission is also allowed, each event resulting in the **creation of one photon** in the EM field at the rate

$$\frac{1}{\tau_{b\to a}} = \frac{2\pi}{\hbar}|\langle a|W_{ab}(\mathbf{r})|b\rangle|^2 \delta[E_a - (E_b - \hbar\omega)], \tag{27.20}$$

which must be **subtracted** because an emission process is a negative contribution to the number of photons absorbed. The above results are

---

[6] Note the symbols carefully since there is a possibility of confusion: $\eta = \sqrt{\mu/\epsilon}$ is the field impedance, which for vacuum is $\sqrt{\mu_0/\epsilon_0} = 377$ Ohms. $n_r = \sqrt{\mu_r \epsilon_r}$ is the dimensionless refractive index which for vacuum is equal to 1, and for a non-magnetic material, $n_r = \sqrt{\epsilon_r}$.

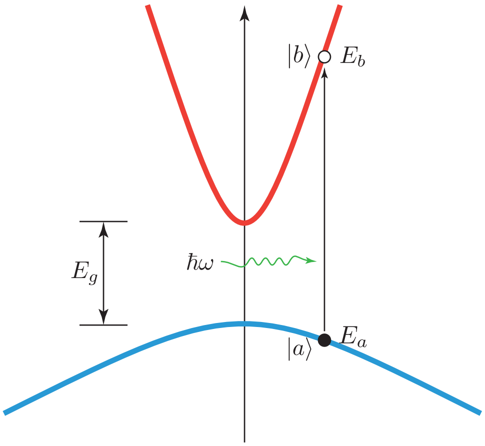


Fig. 27.5 Photon absorption with bandstructure. The absorption process of a single photon by interband transition.

for the single states $|a\rangle$ and $|b\rangle$. A semiconductor crystal has a large number of states in the respective bands, so let's sum the rates for all possible transitions, and divide it by the net volume V to obtain the absorption rate per unit volume (in $\text{s}^{-1}\cdot\text{cm}^{-3}$). Add in the electron spin degeneracy $g_s = 2$ for each $\mathbf{k}$ state, and assume a valley degeneracy $g_v = 1$. The degeneracies must be suitably modified to account for the specific bandstructures of semiconductors[7]. For the absorption process to occur, the lower state $|a\rangle$ has to be occupied (probability $= f_a$) and the higher state $|b\rangle$ has to be empty (probability $= (1 - f_b)$), where f's are the occupation functions. The net absorption rate per unit volume is then given by

$$R_{a\to b} = \frac{2}{V}\sum_{\mathbf{k_a}}\sum_{\mathbf{k_b}}\frac{2\pi}{\hbar}|W_{ba}|^2\delta[E_b - (E_a + \hbar\omega)]f_a(1 - f_b), \qquad (27.21)$$

and the stimulated emission rate (note that this does *not* include the spontaneous emission!) per unit volume is

$$R_{b\to a} = \frac{2}{V}\sum_{\mathbf{k_a}}\sum_{\mathbf{k_b}}\frac{2\pi}{\hbar}|W_{ab}|^2\delta[E_a - (E_b - \hbar\omega)]f_b(1 - f_a). \qquad (27.22)$$

The summation runs over *all* valence band electron states $\mathbf{k_a}$ and conduction band electron states $\mathbf{k_b}$, including those that do not meet the criteria $E_b - E_a = \hbar\omega$. The energy conservation requirement is automatically taken care of by the Dirac delta functions. We note now that the Dirac delta functions are the same for emission and absorption process, because $\delta[+x] = \delta[-x]$, $|W_{ab}| = |W_{ba}|$, and $f_a(1 - f_b) - f_b(1 - f_a) = f_a - f_b$. Therefore, the net photon *absorption* rate per unit volume is the difference

$$\boxed{R_{abs}(\hbar\omega) = R_{a\to b} - R_{b\to a} = \frac{2}{V}\sum_{\mathbf{k_a}}\sum_{\mathbf{k_b}}\frac{2\pi}{\hbar}|W_{ab}|^2\delta[E_b - (E_a + \hbar\omega)] \times (f_a - f_b).} \qquad (27.23)$$

---

[7] Photons carry an angular momentum of ±ħ depending upon their polarization. Therefore, the conservation of angular momentum couples specific spin states. Here we are considering light with photons of mixed polarization. Angular momentum conservation dictates which bands can be involved in the absorption or emission process, thus providing a way to selectively excite say the light hole, heavy hole, or split-off bands because they differ in their net angular momentum.

To evaluate the sum over states, we must first obtain an expression for the matrix element, which is given by the electron-photon perturbation term

$$W_{ab} = \langle b|\frac{q}{m_e}\mathbf{A}\cdot\hat{\mathbf{p}}|a\rangle. \tag{27.24}$$

At this stage, we need to know the wavefunctions corresponding to the band states $|a\rangle$ and $b\rangle$. In principle, we can resort back to the full machinery of the most accurate bandstructures and electron eigenfunctions, say to first principles or the empirical pseudopotential methods of Chapter 13, and compute the most accurate wavefunctions ψ_a and ψ_b, and evaluate the matrix elements and the following optical properties. This can be done, and is also outlined in Exercise 27.3. However, we will not use the full computational machinery here for two reasons, and adopt the effective mass approximation instead. First, because the effective mass approach leads to analytical results that have high value in gaining understanding, and in designing photonic devices. Second, most optical **emission** spectra occurs from states very close to the bottom of the conduction band and the top of the valence band. At these two locations, the effective mass approximation is at its most accurate.

In the effective mass approximation, the electron wavefunction = (envelope function) × (Bloch lattice function). Let us consider bulk semiconductors in 3D. The valence band state wavefunction is then

$$\psi_a(\mathbf{r}) = C(\mathbf{r})u_v(\mathbf{r}) = \underbrace{\frac{e^{i\mathbf{k}_{\mathbf{v}}\cdot\mathbf{r}}}{\sqrt{V}}}_{\text{Envelope } C(\mathbf{r})} \cdot \underbrace{u_v(\mathbf{r})}_{\text{Bloch lattice function}}, \tag{27.25}$$

and the conduction band state wavefunction is

$$\psi_b(\mathbf{r}) = C'(\mathbf{r})u_c(\mathbf{r}) = \frac{e^{i\mathbf{k}_{\mathbf{c}}\cdot\mathbf{r}}}{\sqrt{V}}u_c(\mathbf{r}). \tag{27.26}$$

Since the spatial part of the vector potential for the EM wave from Equation 27.6 is $\mathbf{A} = \hat{e}\frac{A_0}{2}e^{i\mathbf{k}_{\mathbf{op}}\cdot\mathbf{r}}$, the matrix element $W_{ab} = \langle b|\frac{q}{m_e}\mathbf{A}\cdot\hat{\mathbf{p}}|a\rangle$ becomes

$$\begin{aligned} W_{ab} &= \frac{qA_0}{2m_e}\hat{e}\cdot\Big(\int \psi_b^\star e^{i\mathbf{k}_{\mathbf{op}}\cdot\mathbf{r}}\hat{\mathbf{p}}\psi_a d^3\mathbf{r}\Big) \\ &= \frac{qA_0}{2m_e}\hat{e}\cdot\int \Big[\frac{e^{i\mathbf{k}_{\mathbf{c}}\cdot\mathbf{r}}}{\sqrt{V}}u_c(\mathbf{r})\Big]^\star \underbrace{(e^{i\mathbf{k}_{\mathbf{op}}\cdot\mathbf{r}}\hat{\mathbf{p}})}_{\text{operator}}\Big[\frac{e^{i\mathbf{k}_{\mathbf{v}}\cdot\mathbf{r}}}{\sqrt{V}}u_v(\mathbf{r})\Big]d^3\mathbf{r} \\ &= \frac{qA_0}{2m_e}\hat{e}\cdot\int [e^{-i\mathbf{k}_{\mathbf{c}}\cdot\mathbf{r}}u_c^\star(\mathbf{r})] \underbrace{(e^{i\mathbf{k}_{\mathbf{op}}\cdot\mathbf{r}}\hat{\mathbf{p}})}_{\text{operator}}[e^{+i\mathbf{k}_{\mathbf{v}}\cdot\mathbf{r}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{V}, \end{aligned} \tag{27.27}$$

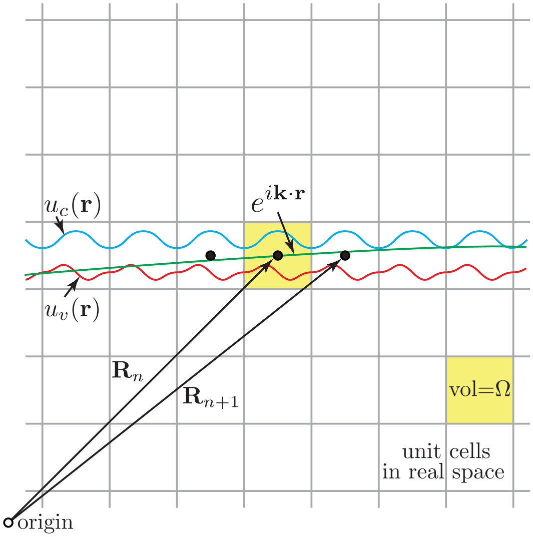


Fig. 27.6 Visualization of the optical matrix element. Explanation of the decomposition of the optical matrix element. Because the matrix element consists of a product of a plane wave part that varies slowly over unit cells, and a part that is periodic in unit cells, the product decomposes into a sum and a cell-periodic integral.

and using $\mathbf{p} = -i\hbar\nabla_r$, we get

$$W_{ab} = \frac{qA_0}{2m_e}\hat{e}\cdot\int[e^{-i\mathbf{k_c}\cdot\mathbf{r}}u_c^\star(\mathbf{r})](e^{i\mathbf{k_{op}}\cdot\mathbf{r}})[e^{+i\mathbf{k_v}\cdot\mathbf{r}}(\hbar\mathbf{k_v}u_v(\mathbf{r}) - i\hbar\nabla u_v(\mathbf{r}))]\frac{d^3\mathbf{r}}{V}$$
$$= \underbrace{\frac{qA_0}{2m_e}\hat{e}\cdot\int e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}[u_c^\star(\mathbf{r})u_v(\mathbf{r})](\hbar\mathbf{k_v})\frac{d^3\mathbf{r}}{V}}_{\text{forbidden}} +$$
$$\underbrace{\frac{qA_0}{2m_e}\hat{e}\cdot\int e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{V}}_{\text{allowed}}. \tag{27.28}$$

The first term is labeled **forbidden** because the value of the integral is $\approx \hbar\mathbf{k}_v\langle\mathbf{k}_c|\mathbf{k}_v\rangle \approx 0$ if we neglect the photon momentum[8]. This is because the states belong to different bands, and are orthogonal. The **allowed** transition matrix element is:

$$W_{ab} = \frac{qA_0}{2m_e}\hat{e}\cdot\int_V e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{V}$$
$$= \frac{qA_0}{2m_e}\hat{e}\cdot\int_V \underbrace{e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}}_{\text{slow}}\underbrace{[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]}_{\text{periodic}}\frac{d^3\mathbf{r}}{\underbrace{N\Omega}_{V}}$$
$$= \frac{qA_0}{2m_e}\hat{e}\cdot\int_V \underbrace{\frac{e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{r}}}{N}}_{\text{slow}}\underbrace{[u_c^\star(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]}_{\text{periodic}}\frac{d^3\mathbf{r}}{\Omega}. \tag{27.29}$$

To visualize the slow and periodic parts inside the integral, refer to Fig. 27.6. The periodic term $u_c(\mathbf{r})\hat{\mathbf{p}}u_c(\mathbf{r})$ repeats in every unit cell in

[8] The "forbidden" absorption cross section is discussed later: though smaller than the allowed transition, it can become important for absorption in indirect bandgap semiconductors such as silicon and germanium. See Exercise 27.8 too.

real space of volume Ω of the crystal. But the slowly varying function of the form $e^{i\mathbf{k}\cdot\mathbf{r}}$ hardly changes inside a unit cell, it changes appreciably only over many many cells. Then, we treat the slowly varying function as a constant inside the unit cell located at $\mathbf{R}_i$, but its value changes slowly from cell to cell. Then, the integral decomposes to

$$W_{ab} = \frac{qA_0}{2m_e}\hat{e}\cdot\underbrace{[\frac{\sum_{n=1}^{N} e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{R_n}}}{N}]}_{\delta_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}} \underbrace{\int_{\Omega}[u_c^{\star}(\mathbf{r})\hat{\mathbf{p}}u_v(\mathbf{r})]\frac{d^3\mathbf{r}}{\Omega}}_{\mathbf{p_{cv}}}. \quad (27.30)$$

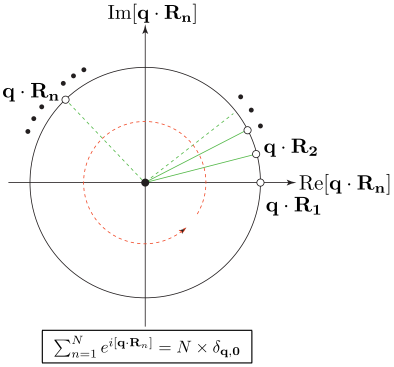

Im[$\mathbf{q}\cdot\mathbf{R_n}$]
$\mathbf{q}\cdot\mathbf{R_n}$
$\mathbf{q}\cdot\mathbf{R_2}$
Re[$\mathbf{q}\cdot\mathbf{R_n}$]
$\mathbf{q}\cdot\mathbf{R_1}$
$\sum_{n=1}^{N} e^{i[\mathbf{q}\cdot\mathbf{R}_n]} = N \times \delta_{\mathbf{q},\mathbf{0}}$

Fig. 27.7 Sum of complex exponentials. The sum of complex exponentials of the form $e^{i\theta_n}$. If the sum is over a large number of phases, the sum $\sum_n e^{i\theta_n}$ is zero, unless $\theta_n = 0$, in which case $\sum_n e^{i\theta_n} = N$. This statement is captured in $\sum_n e^{i\theta_n} = N\delta_{\theta_n,0}$.

The sum runs over all unit cells in real space. Let us visualize the sum $\sum_{n=1}^{N} e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{R_n}}$ of the complex exponential at every unit cell site $\mathbf{R}_n$. First, there are a *lot* of them. Refer to Fig. 27.7 to see why the sum $\frac{\sum_{n=1}^{N} e^{i(-\mathbf{k_c}+\mathbf{k_{op}}+\mathbf{k_v})\cdot\mathbf{R_n}}}{N}$ is zero for all cases except when $-\mathbf{k_c}+\mathbf{k_v}+\mathbf{k_{op}} = 0$, in which case it is evidently unity. The complex numbers $e^{i\theta_n}$ are all on the unit circle on the complex plane, and if there are a *lot* of them, they distribute uniformly around the origin. Thus, their sum tends to have zero real and imaginary parts; further, they are divided by a large number N. But when $\theta_n = 0$, all the points fall at $e^{i0} = 1 + 0i$, and thus the sum is $= N/N = 1$, or unity.

The optical matrix element for **direct, allowed transitions** is thus given by the very important result

$$\boxed{W_{ab} = \frac{qA_0}{2m_e}[\delta_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}](\hat{e}\cdot\mathbf{p_{cv}}).} \quad (27.31)$$

Note that the Kronecker–delta function ensures momentum conservation because $\hbar\mathbf{k_v} + \hbar\mathbf{k_{op}} = \hbar\mathbf{k_c}$. With this form of the optical matrix element, the net absorption rate from Equation 27.23 becomes

$$R(\hbar\omega) = (\frac{qA_0}{2m_e})^2\frac{2}{V}\sum_{\mathbf{k_c}}\sum_{\mathbf{k_v}}\frac{2\pi}{\hbar}|\hat{e}\cdot\mathbf{p_{cv}}|^2 \times$$

$$\delta^2_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}\delta[E_c(\mathbf{k_c}) - (E_v(\mathbf{k_v}) + \hbar\omega)] \times [f_v(\mathbf{k_v}) - f_c(\mathbf{k_c})]. \quad (27.32)$$

Squaring the Kronecker–delta function $\delta^2_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}} = \delta_{\mathbf{k_c},\mathbf{k_v}+\mathbf{k_{op}}}$ leaves it unchanged. We also note at this point that $|\mathbf{k_c}|, |\mathbf{k_v}| >> \mathbf{k_{op}}$. This is because the band-edge states occur around reciprocal lattice vectors $2\pi/a_0$, and the lattice constants $a_0 << \lambda$, the wavelength of light. This is the rationale behind the commonly stated fact: **direct optical transitions are vertical in $E(\mathbf{k})$ diagrams**. Using the Kronecker delta function to reduce the summation over $\mathbf{k}$ states assuming $\mathbf{k_c} = \mathbf{k_v} = \mathbf{k}$, and $\mathbf{k_{op}} \approx 0$, and taking the term $\hat{e}\cdot\mathbf{p_{cv}}$ out of the sum because it does not depend on $\mathbf{k}$, we obtain the net absorption rate per unit volume to be given by the following form, which actually holds also for lower-dimensional structures such as quantum wells, wires, or dots[9]:

$$\boxed{\begin{array}{l} R_{abs}(\hbar\omega) = \frac{2\pi}{\hbar}(\frac{qA_0}{2m_e})^2|\hat{e}\cdot\mathbf{p_{cv}}|^2\times \\ \frac{2}{V}\sum_{\mathbf{k}}\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)] \times [f_v(\mathbf{k}) - f_c(\mathbf{k})]. \end{array}} \quad (27.33)$$

---

[9] Note again that this expression includes the absorption process and the stimulated emission process, but not the spontaneous emission process. For cases where there are a large number of photons involved, this is an excellent approximation. We discuss the detailed effects of spontaneous emission in Chapter 29. Also not that $\mathbf{p}_{cv}$ in Equations 27.31 and 27.33 is the same momentum matrix element encountered in Chapter 12.

27.6 Equilibrium absorption coefficient

We are now ready to evaluate the absorption coefficient. Using the expression for $R_{abs}(\hbar\omega)$ from Equation 27.33, with the photon flux $N_{ph}(\hbar\omega)$ from Equation 27.18, the expression for the absorption coefficient $\alpha(\hbar\omega) = R_{abs}(\hbar\omega)/N_{ph}(\hbar\omega)$ from Equation 27.11 becomes

$$\alpha(\hbar\omega) = \underbrace{(\frac{\pi q^2}{n_r c \epsilon_0 m_e^2 \omega})}_{C_0} |\hat{e} \cdot \mathbf{p_{cv}}|^2 \frac{2}{V} \sum_{\mathbf{k}} \delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)][f_v(\mathbf{k}) - f_c(\mathbf{k})]. \tag{27.34}$$

Notice that the absorption coefficient thus formulated becomes *independent* of the intensity of the incident photon radiation $I \propto A_0^2$ because both $N_{ph}(\hbar\omega) \propto A_0^2$ and $R(\hbar\omega) \propto A_0^2$, and the A_0^2 factor thus cancels in the ratio. This is a signature of a linear process – i.e., the linear absorption coefficient of the semiconductor is a property of the semiconductor alone, and does not dependent on the excitation intensity. With the coefficient $C_0 = (\pi q^2)/(n_r c \epsilon_0 m_e^2 \omega)$ we re-write the absorption coefficient again as the following compact expression which also holds for lower-dimensional structures such as quantum wells, wires, or dots (see Chapter 14):

$$\boxed{\alpha(\hbar\omega) = C_0 \frac{2}{V} \sum_{\mathbf{k}} |\hat{e} \cdot \mathbf{p_{cv}}|^2 \delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)] \times [f_v(\mathbf{k}) - f_c(\mathbf{k})]} \tag{27.35}$$

To evaluate the $\mathbf{k}$-sum, we need to identify the occupation functions $f_v(\mathbf{k})$ and $f_c(\mathbf{k})$. If the semiconductor is in thermal equilibrium, there is one Fermi level E_F, and the occupation is given by the Fermi–Dirac function $f(E) = (1 + \exp[(E - E_F)/k_B T])^{-1}$ at temperature T. When the semiconductor is pushed to non-equilibrium by either optical excitation or electrical injection of excess carriers[10], the occupation functions are conveniently modeled by retaining the Fermi–Dirac form. But the single Fermi-energy E_F splits to two quasi-Fermi levels: one for electrons in the conduction band F_c, and the other for electrons in the valence band F_v. The occupation functions are then given by

$$f_v(\mathbf{k}) = \frac{1}{1 + \exp\left(\frac{E_v(\mathbf{k}) - F_v}{k_b T}\right)}, \text{ and}$$
$$f_c(\mathbf{k}) = \frac{1}{1 + \exp\left(\frac{E_c(\mathbf{k}) - F_c}{k_b T}\right)}. \tag{27.36}$$

We will consider non-equilibrium conditions in Chapter 29. Under thermal equilibrium, $F_c = F_v = E_F$, and there is only one Fermi level. For an undoped semiconductor, E_F locates close to the middle of the bandgap[11]. Then, as $T \to 0$ K, $f_v(\mathbf{k}) \to 1$ and $f_c(\mathbf{k}) \to 0$. Actually, these conditions hold fine even at room temperature for wide-bandgap semiconductors with little error. Converting the sum in Equation 27.35

[10] This was discussed in earlier chapters, particularly in Chapter 18 for pn junctions.

[11] When photons are incident on the semiconductor, it is by definition not in equilibrium and $F_c \neq F_v$. But we assume that the intensity of the EM wave is low enough to ensure that $F_c \approx F_v \approx E_F$.

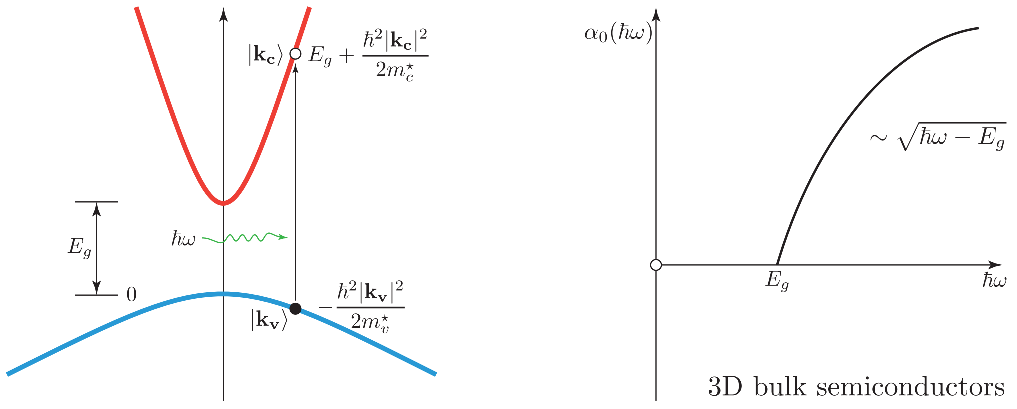

$|\mathbf{k_c}\rangle$
$E_g + \frac{\hbar^2|\mathbf{k_c}|^2}{2m_c^\star}$
E_g
$\hbar\omega$
0
$|\mathbf{k_v}\rangle$
$-\frac{\hbar^2|\mathbf{k_v}|^2}{2m_v^\star}$
$\alpha_0(\hbar\omega)$
$\sim \sqrt{\hbar\omega - E_g}$
E_g
$\hbar\omega$
3D bulk semiconductors

Fig. 27.8 Absorption spectrum of bulk semiconductors. Definition of various energies, and the equilibrium absorption spectrum of bulk (3D) semiconductors $\alpha_0(\hbar\omega)$.

to an integral using the usual prescription, we get the **equilibrium absorption coefficient** to be

$$\boxed{\alpha_0(\hbar\omega) = C_0|\hat{e}\cdot\mathbf{p_{cv}}|^2\frac{2}{V}\times\int_{\mathbf{k}}\frac{d^3\mathbf{k}}{\frac{(2\pi)^3}{V}}\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)].} \quad (27.37)$$

Note that the volume term V cancels. From Fig. 27.8, the optical transition can occur only for **k** states that satisfy

$$\begin{aligned} E_c(\mathbf{k}) &= E_g + \frac{\hbar^2k^2}{2m_e^\star} \\ E_v(\mathbf{k}) &= -\frac{\hbar^2k^2}{2m_h^\star} \\ \hbar\omega = E_c(\mathbf{k}) - E_v(\mathbf{k}) &= E_g + \frac{\hbar^2k^2}{2m_r^\star} \\ \frac{1}{m_r^\star} &= \frac{1}{m_e^\star} + \frac{1}{m_h^\star}. \end{aligned} \quad (27.38)$$

Using spherical coordinates in the 3D **k**-space, $d^3\mathbf{k} = k^2\sin\theta dkd\theta d\phi$, we convert the variables from wavevector to energy. Assuming $E = \frac{\hbar^2k^2}{2m_r^\star}$, we break up $k^2\sin\theta dkd\theta d\phi$ into three parts: $k^2\cdot dk = (\frac{2m_r^\star}{\hbar^2})E\cdot\frac{1}{2}(\frac{2m_r^\star}{\hbar^2})^{\frac{1}{2}}\frac{dE}{\sqrt{E}} = \frac{1}{2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{E}dE$, the second part being $\sin\theta d\theta$ and the third part $d\phi$. When we integrate over all k-space, the angular parts evaluate to $\int_0^\pi\sin\theta d\theta = 2$ and $\int_0^{2\pi}d\phi = 2\pi$.

The absorption coefficient then becomes

$$\alpha_0(\hbar\omega) = C_0|\hat{e}\cdot\mathbf{p_{cv}}|^2\frac{2}{(2\pi)^3}\cdot(2\pi)\cdot(2)\cdot\frac{1}{2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}} \times \underbrace{\int_0^\infty dE\sqrt{E}\times\delta[E-(\hbar\omega-E_g)]}_{\sqrt{\hbar\omega-E_g}}, \tag{27.39}$$

which reduces to

$$\boxed{\alpha_0(\hbar\omega) = C_0|\hat{e}\cdot\mathbf{p_{cv}}|^2\underbrace{\frac{2}{(2\pi)^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{\hbar\omega-E_g}}_{\rho_J(\hbar\omega-E_g)},} \tag{27.40}$$

where we have defined the optical joint density of states function for bulk 3D semiconductors as[12]

$$\rho_J(u) = \frac{g_s}{(2\pi)^2}\cdot(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\cdot\sqrt{u}. \tag{27.41}$$

Fig. 27.8 shows the equilibrium absorption spectrum $\alpha_0(\hbar\omega)$ of a bulk 3D semiconductor. Using typical values of effective masses and material constants, it may be verified that the interband absorption coefficient for GaN for example is of the order of $\sim 10^5$ cm^{-1}, as indicated in Fig. 27.4 at the beginning of this chapter. The absorption coefficient is zero for photon energies below the bandgap of the semiconductor, as is intuitively expected[13].

Instead of leaving the expression for the absorption coefficient in terms of the parameter C_0, we use the fundamental Rydberg energy $R_\infty = \frac{q^2}{4\pi\epsilon_0(2a_B)}$, the Bohr radius $a_B = \frac{\hbar}{m_0 c\alpha}$, and the fine structure constant $\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c}$ to write the absorption coefficient as

$$\boxed{\alpha_0(\hbar\omega) = (\frac{4\pi^2\alpha}{n_r})\cdot(R_\infty a_B^2)\cdot(\frac{\frac{2|\hat{e}\cdot\mathbf{p_{cv}}|^2}{m_0}}{\hbar\omega})\cdot\rho_r(\hbar\omega-E_g),} \tag{27.42}$$

where we have split off the dimensionless term $2|\hat{e}\cdot\mathbf{p_{cv}}|^2/m_0\hbar\omega = E_P/(\hbar\omega)$ called the **oscillator strength**[14]. Note that as discussed in Chapter 12 on the $\mathbf{k}\cdot\mathbf{p}$ bandstructure theory of semiconductors, the rough order of $E_P = 2|\hat{e}\cdot\mathbf{p_{cv}}|^2/m_0 \approx 20$ eV for most bulk semiconductors. The coefficients decompose to reveal a proportionality to the fine-structure constant. The term $R_\infty a_B^2$ has units eV.cm^2, and the reduced density of states is in 1/eV.cm^3, which leads to the net units cm^{-1}. Let us examine how this band-edge absorption coefficient changes in lower dimensional quantum structures.

27.7 Quantum wells, wires, and dots

The derivation of the equilibrium absorption coefficient expression of Equation 27.37 is general, and therefore this experession applies to

[12] Note that the optical joint DOS $\rho_J(...)$ takes the same functional form as the d-dimensional DOS for parabolic bands, albeit with a reduced effective mass and the excess energy $\hbar\omega - E_g$ as the argument.

[13] This is true in the absence of excitons and defects. These exceptions are discussed later in this chapter.

[14] The reason for this name will be discussed shortly.

lower dimensional quantized structures. Notice that because equilibrium absorption coefficient holds for a filled valence band with $f_v(\mathbf{k}) = 1$ and an empty conduction band $f_c(\mathbf{k}) = 0$, it can be written by splitting off the **optical joint density of states** $\rho_J(\hbar\omega)$:

$$\alpha_0(\hbar\omega) = \frac{\pi q^2 |\hat{e} \cdot \mathbf{p}_{cv}|^2}{n_r c \epsilon_0 m_e^2 \omega} \rho_J(\hbar\omega),$$

$$\text{where } \boxed{\rho_J(\hbar\omega) = \frac{2}{V} \sum_{\mathbf{k}} \delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)].} \qquad (27.43)$$

The units of $\rho_J(\hbar\omega)$ are $1/(\text{energy} \cdot \text{volume})$, which is clearly an energy density of states. The advantage of writing the absorption coefficient in this way is that $\rho_J(\hbar\omega)$ depends exclusively on the equilibrium properties of the semiconductor bandstructure which are well defined. In the next section, we will see how $\rho_J(\hbar\omega)$ defines *critical points* in the bandstructure far from the band edges. In this section, we explore the band-edge $\rho_J(\hbar\omega)$ and the resulting $\alpha_0(\hbar\omega)$ in low dimensions.

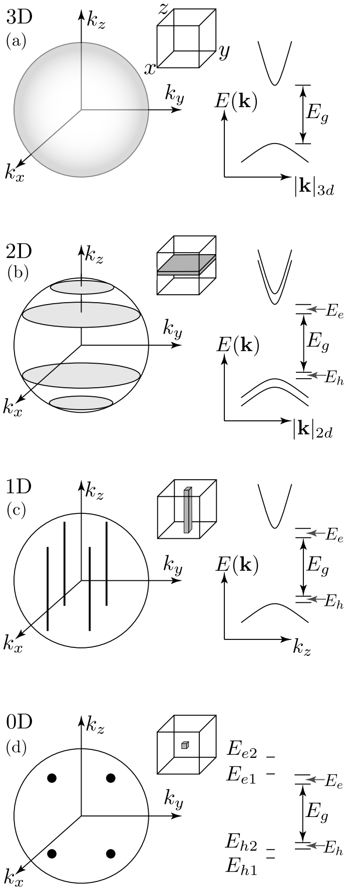


Fig. 27.9 Energy eigenvalues pictured in the **k**-space in (a) 3D bulk and in low dimensions: (b) 2D, (c) 1D, and (d) 0D.

Let us explore the physical meaning of $\rho_J(\hbar\omega)$. How many electron k states are available per unit volume to emit, or absorb photons in the energy interval $[h\omega, h(\omega + d\omega)]$? The answer to this question is $\rho_J(\hbar\omega)d(\hbar\omega)$. **Therefore the optical JDOS quantifies the strength of the light-matter interaction.** As was introduced in the sneak peek in Chapter 16, it captures the property of light through $\hbar\omega$, and matter through the semiconductor parameters the bandgap E_g, and electron energy eigenvalues $E_c(\mathbf{k})$ and $E_v(\mathbf{k})$. Let us now evaluate $\rho_J(\hbar\omega)$ for near band-edge transitions in optically active semiconductor electron states are unconfined and free to move in 3D bulk, or confined in quantum wells (2D), quantum-wires (1D), or quantum dots (0D)[15].

To count the number of electron states in 3D, assume the electrons are confined in a cubic box of size $L_x = L_y = L_z = L$ as shown in Fig. 27.9. In the sneak peek of Chapter 16, we had used hard-wall boundary conditions. Here we use periodic boundary conditions to show the invariance of the results. Periodic boundary conditions on the electron wavefunctions for the box restrict the values of wavevectors to $\mathbf{k} = (k_x, k_y, k_z) = \frac{2\pi}{L}(n_x, n_y, n_z)$, where (n_x, n_y, n_z) is a triplet of integers each of whose values can be $0, \pm 1, \pm 2, ...$ This defines a 3D lattice in the $\mathbf{k} = (k_x, k_y, k_z)$ space as shown in Fig. 27.9 (a). Each k-point denotes an allowed electron state, occupying a volume $(\frac{2\pi}{L})^3 = \frac{(2\pi)^3}{V}$. Because the volume of the semiconductor cube is $V = L^3$ and we are interested in the JDOS $\rho_J(\hbar\omega)$ *per unit volume*, we also define the *total* JDOS as $D_J(\hbar\omega)$ which is related to the total JDOS by $\rho_J(\hbar\omega) = D_J(\hbar\omega)/V$.

3D bulk: The JDOS $\rho_J(k)$ in the **k**-space counts the same states as in the energy space, meaning $\rho_J(k)dk = D_J(\hbar\omega)d(\hbar\omega)$. For **allowed transitions** in a **direct-bandgap semiconductor**, using Equation 27.38

[15] The quantum confined structures are routinely used in practical semiconductor LEDs, SOAs, modulators, and lasers as discussed in Chapters 28 and 29.

in the form $\frac{\hbar^2}{2m_r^\star}(k_x^2+k_y^2+k_z^2)=\hbar\omega-E_g$, we obtain

$$\rho_J^{3d}(k)d^3k = 2\cdot\frac{4\pi k^2 dk}{\frac{2\pi}{L_x}\frac{2\pi}{L_y}\frac{2\pi}{L_z}} = D_J(\hbar\omega)d(\hbar\omega)$$

$$\implies \boxed{\rho_J^{3d}(\hbar\omega)=\frac{1}{2\pi^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{\hbar\omega-E_g},} \quad (27.44)$$

where $\rho_J^{3d}(\hbar\omega)=D_J(\hbar\omega)/V$ was used. The typical semiconductor units of $\rho_J^{3d}(\hbar\omega)$ are in $\frac{1}{\mathrm{eV\cdot cm^3}}$ as may be verified from Equation 27.44. Fig. 27.10 shows the $\rho_J^{3d}(\hbar\omega)\propto\sqrt{\hbar\omega-E_g}$. This implies the corresponding $\alpha_0(\hbar\omega)=\frac{\pi q^2|\hat{e}\cdot\mathbf{p}_{cv}|^2}{\eta c\epsilon_0 m_e^2\omega}\rho_J(\hbar\omega)\propto\frac{\sqrt{\hbar\omega-E_g}}{\hbar\omega}$. For this reason, a plot of the measured $(\alpha_0(\hbar\omega)\hbar\omega)^2$ as a function of the photon energy $\hbar\omega$ yields E_g as the x-axis intercept, and the slope can be used to extract $\mathbf{p}_{cv}$. Such **Tauc** plots are used to measure E_g and $\mathbf{p}_{cv}$.

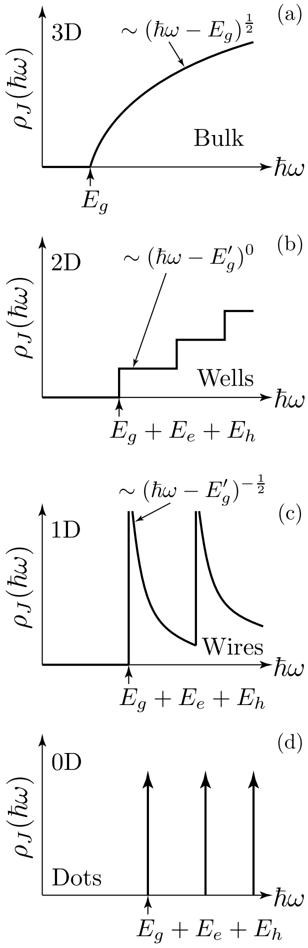


Fig. 27.10 Optical joint density of states for 3D to 0D semiconductor structures for parabolic conduction and valence bands. Note that the absorption coefficient is related to the JDOS via Equation 27.43.

2D quantum wells: The JDOS for quantum confined electron states is found in the same manner as for the 3D in the corresponding lower-dimensional k. As shown in Fig. 27.9 (b), in quantum wells used in LEDs and lasers, the electron and hole states are free to move in two dimensions (x,y) because of large dimensions $L_x=L_y=L$ and confined in the third z-direction because of a heterostructure quantum well confining potential to a length L_z, smaller than the electron deBroglie wavelength. This situation is realized if we create a structure in which $L_z << L$, where L_z is a few nm, whereas L is the size of the wafer – in a few 100s of μm or in mm. For such a quantum-confined structure, Equation 27.38 now takes the form

$$\frac{\hbar^2}{2m_r^\star}(\underbrace{k_x^2+k_y^2}_{k_\parallel^2}+k_z^2)=\hbar\omega-E_g$$

$$\implies \frac{\hbar^2}{2m_r^\star}k_\parallel^2=\hbar\omega-[E_g+\underbrace{\frac{\hbar^2}{2m_e^\star}(\frac{\pi}{L_z}n_z)^2}_{E_e}+\underbrace{\frac{\hbar^2}{2m_h^\star}(\frac{\pi}{L_z}n_z)^2}_{E_h}], \quad (27.45)$$

where we identify the blue-shift compared to E_g in the minimum transition energy due to quantum confinement in the conduction and the valence bands E_e+E_h. This quantum-engineering of the effective bandgap is a very powerful tool for engineering light-matter interactions in semiconductor heterostructures.

The allowed electron modes then form a dense 2D lattice in the $(k_x,k_y)=\frac{2\pi}{L}(n_x,n_y)$ planes, separated by large $k_z=\frac{2\pi}{L_z}n_z$ in the k_z-

direction. The band-edge JDOS in 2D is then

$$\rho_J^{2d}(k)d^2k = 2\cdot\frac{(2\pi k_{\|}dk_{\|})\frac{2\pi}{L_z}}{\frac{2\pi}{L_x}\frac{2\pi}{L_y}\frac{2\pi}{L_z}} = D_J(\hbar\omega)d(\hbar\omega)$$

$$\implies \boxed{\rho_J^{2d}(\hbar\omega) = \frac{m_r^\star}{\pi\hbar^2 L_z}\Theta[\hbar\omega-(E_g+E_n+E_p)],} \qquad (27.46)$$

where we again used $\rho_J^{2d}(\hbar\omega) = D_J(\hbar\omega)/V$. The Heaviside unit-step function $\Theta[...]$ in the 2D optical JDOS yet again reinforces our intuition that the smallest photon energy that can interact with the quantum well electron states must have an energy $h\nu = E_g + E_n + E_p$ because quantum confinement has pushed the minimum energy in the conduction band up to $E_c + E_n$ and in the valence band down to $E_v - E_p$. The units of $\rho_J^{2d}(\hbar\omega)$ is still $\frac{1}{\text{eV}\cdot\text{cm}^3}$, the same as for 3D – this is verified from Equation 27.46. Steps appear in the JDOS at higher subband energies for electrons and holes as seen in Fig. 27.10 (b).

1D quantum wires: Continuing to lower dimensions, it is left to the reader to show that for electron states in 1D quantum wires confined in the (x, y) plane to (L_x, L_y) and free to move in the z-direction, the optical JDOS is

$$\boxed{\rho_J^{1d}(\omega) = \frac{2}{\pi L_x L_y}\sqrt{\frac{2m_r^\star}{\hbar^2}}\frac{1}{\sqrt{\hbar\omega-(E_g+E_n+E_p)}},} \qquad (27.47)$$

where the quantum confinement energies are now $E_n = \frac{\hbar^2}{2m_e^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2]$ and $E_p = \frac{\hbar^2}{2m_h^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2]$, shown in Fig. 27.10 (c).

0**D quantum dots:** For 0D quantum dots in which electrons are confined in all three directions to boxes of size (L_x, L_y, L_z), the structure mimics an atomic system, giving an optical JDOS

$$\boxed{\rho_J^{0d}(\omega) = \frac{2}{L_x L_y L_z}\delta[\hbar\omega-(E_g+E_n+E_p)],} \qquad (27.48)$$

where $\delta[...]$ is the Dirac delta function highlighting that the electron energies are discrete and atomic-like, and do not form extended states or bands as shown in Fig. 27.10 (d). The quantum confinement energies are now $E_n = \frac{\hbar^2}{2m_e^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2 + (\frac{\pi}{L_z}n_z)^2]$ and $E_p = \frac{\hbar^2}{2m_h^\star}[(\frac{\pi}{L_x}n_x)^2 + (\frac{\pi}{L_y}n_y)^2 + (\frac{\pi}{L_z}n_z)^2]$. It may be verified again that the units of $\rho_J^{0d}(\nu)$ is exactly the same as in other dimensions.

27.8 Critical points

Fig. 27.3 (b) showed the absorption spectrum of silicon at photon energies not just near the band-edge, but well above it. Several sharp peaks are seen for photon energies well above the band-edge at energies called **critical points**. These features provide experimental evidence of the full bandstructure of the semiconductor far from the band edges. In this section, we discuss the nature of critical points, their classification, and their consequences in the optical spectra.

The absorption coefficient $\alpha_0(\hbar\omega)$ and the optical JDOS $\rho_J(\hbar\omega)$ are related to the imaginary part of the frequency-dependent dielectric constant $\epsilon_2(\omega)$ by the useful relations

$$\boxed{\epsilon_2(\hbar\omega) = \frac{\pi q^2 |\hat{e}\cdot\mathbf{p}_{cv}|^2}{\epsilon_0 m_e^2 \omega^2}\rho_J(\hbar\omega) = \frac{\eta c}{\omega}\alpha_0(\hbar\omega).} \tag{27.49}$$

Using the Kramers–Kronig relation (Equation 26.13), the frequency dependent dielectric constant $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega)$ is given by

$$\boxed{\epsilon_2(\omega) = \frac{\pi q^2}{\epsilon_0 m^2 \omega^2}\cdot\frac{2}{V}\sum_{\mathbf{k}}|\hat{e}\cdot\mathbf{p}_{cv}|^2\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)], \text{ and}}$$
$$\boxed{\epsilon_1(\omega) = 1 + \frac{q^2}{\epsilon_0 m}\cdot\frac{2}{V}\sum_{\mathbf{k}}\frac{2}{m_e\hbar\omega_{cv}(\mathbf{k})}\frac{|\hat{e}\cdot\mathbf{p}_{cv}|^2}{\omega_{cv}^2(\mathbf{k}) - \omega^2},} \tag{27.50}$$

where $\hbar\omega_{cv}(\mathbf{k}) = E_c(\mathbf{k}) - E_v(\mathbf{k})$. Comparing the above microscopic expression of the real part of the dielectric constant to the classical Drude–Lorentz oscillator model discussed in Chapter 26 Equation 26.12 given by $\epsilon_1(\omega) = 1 + \sum_j \frac{N_j q^2}{\epsilon_0 m_e}\frac{1}{\omega_j^2 - \omega^2}$, we find the quantum result is equivalent to the classical one for no damping, if the dimensionless quantity $\frac{2|\hat{e}\cdot\mathbf{p}_{cv}|^2/m_e}{\hbar\omega_{cv}} = \frac{1}{2}VN_j$, which is half the total number of oscillators of type j in the classical model[16]. The dimensionless quantity $f_{cv} = \frac{2|\hat{e}\cdot\mathbf{p}_{cv}|^2/m_e}{\hbar\omega_{cv}}$ is thus called the **oscillator strength**.

Though our earlier discussions have been limited to near band-edge energies of the conduction and valence band, Equation 27.50 provides the recipe to compute the real and imaginary parts of the **dielectric constant at all frequencies if the full bandstructure of the semiconductor is known**. Because by optical spectroscopy we can experimentally measure either $\epsilon_1(\omega)$ or $\epsilon_2(\omega)$, a comparison of the experimental and computed $\epsilon_1(\omega)$ serve as a stringent test of the bandstructure model itself. Instead of discussing the complete evaluation, we focus on the critical points using simplified bandstructure models to understand the principles of the analysis.

Equation 27.49 indicates that $\epsilon_2(\omega)$ is directly related to the optical JDOS $\rho_J(\hbar\omega)$, which was already shown in Fig. 27.10. Since the JDOS is an integral of a Dirac delta function, we can use the identity $\delta[f(x)] = \sum_i \frac{\delta[x - x_i]}{|f'(x_i)|}$ for all zeroes x_i of the argument of the Dirac delta

[16] The factor of 2 due to electron spin degeneracy has no classical analog. Also note that $\hbar\omega_{cv} = E_c - E_v$, which for direct-gap semiconductors is the bandgap E_g.

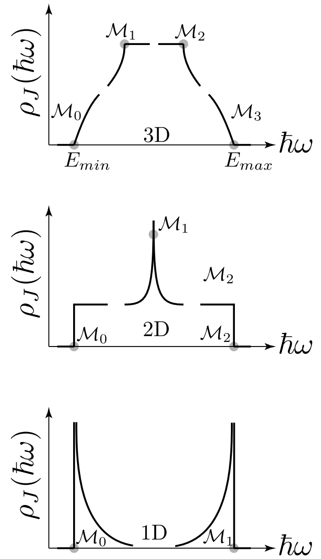


Fig. 27.11 Critical points in the optical joint density of states.

Table 27.1 Dependence of $\rho_J(\hbar\omega)$ on the energy E around critical points E_0.

Type	$\rho_J(E < E_0) \sim$	$\rho_J(E > E_0) \sim$
3D		
$\mathcal{M}_0$	0	$\sqrt{E - E_0}$
$\mathcal{M}_1$	$C - \sqrt{E - E_0}$	C
$\mathcal{M}_2$	C	$C - \sqrt{E - E_0}$
$\mathcal{M}_3$	$\sqrt{E - E_0}$	0
2D		
$\mathcal{M}_0$	0	C
$\mathcal{M}_1$	$\ln(\frac{1}{E_0 - E})$	$\ln(\frac{1}{E - E_0})$
$\mathcal{M}_2$	C	0
1D		
$\mathcal{M}_0$	0	$\frac{1}{\sqrt{E - E_0}}$
$\mathcal{M}_1$	$\frac{1}{\sqrt{E_0 - E}}$	0

function that make $f(x_i) = 0$. The JDOS may then be re-expressed as

$$\rho_J(\hbar\omega) = \frac{2}{V}\sum_{\mathbf{k}} \delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)]$$

$$= \frac{2}{V}\int \frac{d^3k}{\frac{(2\pi)^3}{V}} \delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)] = \frac{1}{4\pi^3}\int \sum_{k_\omega} \frac{\delta[k - k_\omega]d^3k}{\nabla_{\mathbf{k}}(E_c(\mathbf{k}) - E_v(\mathbf{k}))}$$

$$\implies \rho_J(\hbar\omega) = \frac{1}{4\pi^3}\int_{E_{cv}(\mathbf{k})=\hbar\omega} \frac{dS_k}{\nabla_{\mathbf{k}}(E_c(\mathbf{k}) - E_v(\mathbf{k}))}. \tag{27.51}$$

In this reformulation, k_ω indicates those **k**-points in the Brillouin zone that satisfy energy conservation in the absorption of the photon via the relation $E_c(\mathbf{k}) - E_v(\mathbf{k}) = E_{cv}(\mathbf{k}) = \hbar\omega$. The set of these points form a surface S_k in the **k**-space. It is clear then that $\rho_J(\hbar\omega)$ will have **singularities** at those **k**-points for which the denominator $\nabla_{\mathbf{k}}(E_c(\mathbf{k}) - E_v(\mathbf{k})) = 0$. These are precisely the **critical points** in the **k**-space. They occur when (a) $\nabla_{\mathbf{k}} E_c(\mathbf{k}) = \nabla_{\mathbf{k}} E_v(\mathbf{k}) = 0$, or (b) $\nabla_{\mathbf{k}} E_c(\mathbf{k}) = \nabla_{\mathbf{k}} E_v(\mathbf{k}) \neq 0$. The first set of points (a) occur when the group velocity of the states are *simultaneously* zero, which occurs if band minima and maxima occur at the same **k**-points. Clearly this is the case at the Γ-point band edge of all direct-gap semiconductors. The second set of points (b) occur at the **k**-points for which the conduction and valence band group velocities match, and the bands are *parallel* in the $E(\mathbf{k})$ diagram, such that $E_c(\mathbf{k}) - \mathbf{E_v}(\mathbf{k})$ is independent of **k**.

As a result of the denominator in Equation 27.51 going to zero, the integral over the diverging function results in sharp features with discontinuous derivatives in the JDOS at the corresponding energies. These are called **Van Hove singularities** after the physicist Leon Van Hove who pointed out their significance (he first discussed them in the context of the *phonon* DOS). Fig. 27.11 indicates the critical points labeled $\mathcal{M}_0$ to $\mathcal{M}_3$ and the characteristic features of $\rho_J(\hbar\omega)$ in 3D, 2D, and 1D at energies E_0. Table 27.1 provides their functional forms. The shapes of the critical points at the lower band extremum $E = E_{min}$ is identical to what we have already evaluated and shown in Fig. 27.10, and the features at the upper band extremum $E = E_{max}$ are their mirror reflections. The shapes of the Van Hove singularities inside the bands are obtained by expanding the energy around the critical point:

$$E_{cv}(k) = E_0 + \alpha_1 k_1^2 + \alpha_2 k_2^2 + \alpha_3 k_3^2 + \mathcal{O}(> k^2) \tag{27.52}$$

They depend on the sign of α_i which are the reduced effective masses of the kind $\alpha_1 = \frac{\hbar^2}{2}(\frac{1}{m_c^\star(k_1)} + \frac{1}{m_v^\star(k_1)})$. Unlike at the bottom of the conduction band and top of the valence band, the effective masses deep in the bands may be negative. Looking back at Fig. 27.3 (b), since $\alpha_0(\hbar\omega) \propto \rho_J(\hbar\omega)$, the sharp peaks at such as E_1 and E_2 can be seen also in 3D. This is because the constant energy "surface" in **k**-space can be a *plane* (2D), or a *line* (1D). In the bandstructure the lowest conduction band and highest valence bands are indeed seen to be nearly parallel around the line labeled E_1 in the bandstructure, for which the condition $\nabla_{\mathbf{k}} E_c(\mathbf{k}) = \nabla_{\mathbf{k}} E_v(\mathbf{k}) \neq 0$ is satisfied. Exercise 27.7 investigates the shapes of Van Hove singularities further.

For the simple case of the 3D, from Equation 27.44 we know that $\rho_J^{3d}(\hbar\omega) = \frac{1}{2\pi^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{\hbar\omega - E_g}$ for $\hbar\omega > E_g$ and $= 0$ for $\hbar\omega < E_g$. Using Equation 27.49, we realize that the complex part of the dielectric constant $\epsilon_2(\omega)$ is now known for all ω near E_g. Then, using Equation 27.50, the real part $\epsilon_1(\omega)$ is also obtained using the Kramers–Kronig transformation. Denoting $x = \hbar\omega/E_g$, the real and imaginary parts of the dielectric function are obtained as

$$\epsilon_2(\omega) = \underbrace{\frac{\sqrt{2}q^2(m_r^\star)^{\frac{3}{2}}|\hat{e}\cdot\mathbf{p}_{cv}|^2}{\pi\hbar\epsilon_0 m_e^2 E_g^{\frac{3}{2}}}}_{C}\frac{\sqrt{x-1}}{x^2}\Theta[x-1] = C\frac{\sqrt{x-1}}{x^2}\Theta[x-1], \text{ and}$$

$$\epsilon_1(\omega) = B + C\frac{2-\sqrt{1+x}}{x^2}\Theta[x-1] + C\frac{2-\sqrt{1+x}-\sqrt{1-x}}{x^2}\Theta[1-x]. \tag{27.53}$$

Here B is a constant that depends on the semiconductor properties. Fig. 27.12 (a) shows a plot of the dependence of the real and imaginary parts of the dielectric constants given by Equation 27.53. This behavior is very close to what is measured experimentally. Fig. 27.12 (b) shows the corresponding **Tauc plot** of the squared product of the photon energy and absorption coefficient. The intercept yields the bandgap, and from the slope the unknown quantities in the constant C can be obtained experimentally.

We have thus derived the interband absorption coefficients of semiconductors at equilibrium. Many interesting photonic effects occur when the semiconductor is pushed out of equilibrium. In this chapter we will see how the absorption coefficient is modified by electric or magnetic fields in Section 27.11. We will discuss the uses of non-equilibrium absorption in solar cells and photodetectors in Chapter 28. In Chapter 29, we will see how electrically driving semiconductor junctions weakly out of equilibrium leads to **electroluminescence** which forms the basis of LEDs, and how special non-equilibrium conditions leads to the dramatic effect of **negative optical absorption coefficients**, which provide **optical gain** and are the basis of semiconductor lasers. All these phenomena and device applications are intricately tied to the specific nature of the equilibrium semiconductor absorption spectra, and are sometimes affected by features other than the band-edge absorption. So in the rest of this chapter, we investigate additional light-matter interaction processes (other than the interband transitions of this section) that add features to the equilibrium absorption spectra such as forbidden and indirect transitions, intersubband transitions, and absorption of light by free carriers, impurities, and excitons that were qualitatively introduced in Fig. 27.4.

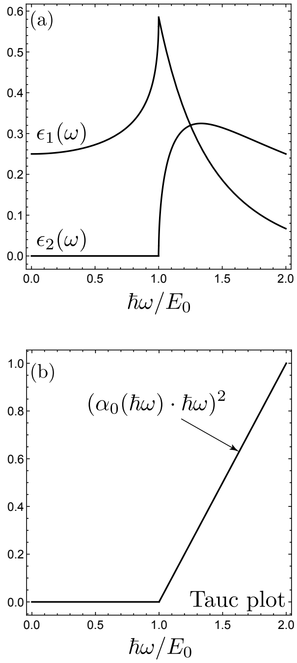


Fig. 27.12 (a) Real and imaginary parts of the frequency dependent dielectric constant in relative values for the $\mathcal{M}_0$ critical point in 3D. (b) The square of the quantity $\alpha_0(\hbar\omega)\cdot(\hbar\omega)$ from part (a) intersects the y-axis at the critical point E_0, which for 3D $\mathcal{M}_0$ point is the energy bandgap of the 3D semiconductor.

27.9 Forbidden and indirect absorption

Forbidden transitions: Consider again Equation 27.28 for transition of the electron state upon photon absorption. Though momentum conservation *allows* only vertical (or direct) transitions, a vertical transition may actually be forbidden by selection rules if the matrix element is zero[17]. This occurs if the valence and conduction state wavefunctions are of the same parity leading to $\langle \mathbf{k}_c | \hat{\mathbf{p}} | \mathbf{k}_v \rangle = 0$. The term labeled "forbidden" is small, but may not be exactly zero. Following the same procedure of summing over all $\mathbf{k}$ states, in Exercise 27.8 it is shown that the absorption coefficient for the forbidden direct transition is

$$\alpha_0(\hbar\omega) = \frac{q^2 (2m_r^\star)^{\frac{5}{2}} f'}{6\pi\epsilon_0 n_r c m_e^2 \hbar^2} \frac{(\hbar\omega - E_g)^{\frac{3}{2}}}{\hbar\omega}, \qquad (27.54)$$

for 3D carriers. Here $f' = |\int d^3 r u^\star_{v,\mathbf{k}'}(\mathbf{r}) u_{c,\mathbf{k}}(\mathbf{r})|^2$ takes a value $0 \leq f' \ll 1$. Note that the dependence on the excess photon energy in 3D goes as $(\hbar\omega - E_g)^{3/2}$ instead of the square root dependence for the direct allowed transition.

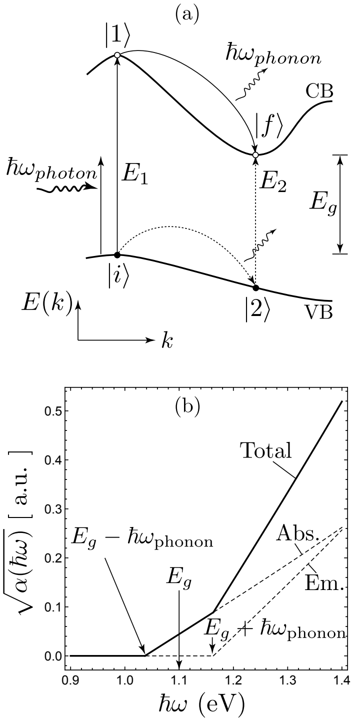


Fig. 27.13 (a) Interband photon absorption via virtual states for indirect bandgap semiconductors, and (b) the resulting absorption spectrum.

Indirect transitions: The stringent requirement of momentum conservation allows only transitions that are vertical in $\mathbf{k}$ in electron-photon interactions due to the very small momentum of photons. The requirement of momentum conservation cannot be violated in any process. However, the requirement of energy conservation can be "violated" for *a part of* a process, as long as it is *rigidly conserved for the total process*. This wiggle room is necessary to explain why indirect bandgap semiconductors such as silicon or germanium have an appreciable absorption coefficient at and even below the bandgap E_g. Though the indirect absorption coefficient is small $\alpha \sim 10^3$ cm^{-1} compared to the direct $\alpha \sim 10^5$ cm^{-1}, it is of immense practical value in photodetectors and solar cells that are discussed in Chapter 28.

Fig. 27.13 (a) shows an indirect absorption process, and Fig. 27.13 (b) shows the resulting absorption spectrum, which is close to what is observed experimentally. The indirect bandstructure of silicon along the $\Gamma - X$, or (111) direction in $\mathbf{k}$-space is a good example for Fig. 27.13 (a). The fundamental (indirect) bandgap is E_g. An electron cannot make a transition directly from state $|i\rangle$ to $|f\rangle$ because it violates momentum conservation. A third particle is needed that can supply the necessary "kick" in momentum for this transition. As was discussed in Chapter 22, phonons can supply momentum of $\hbar\mathbf{k}_{\text{phonon}}$ where $0 \leq |\mathbf{k}_{\text{phonon}}| \leq \pi/a$ spans the Brillouin zone, but energies of $0 \leq \hbar\omega_{\text{phonon}} \leq \hbar\omega_{LO}$ less than the maximum LO phonon energy.

The indirect absorption process therefore involves *three* particles: the electron, the photon, and a phonon. The transition may be pictured to proceed via the **virtual states** $|1\rangle$ and $|2\rangle$ in the following manner. The initial state of the three-particle system is $|i\rangle|N_{\text{photon}}\rangle|N_{\text{phonon}}\rangle$, and the final state for the transition is $|f\rangle|N_{\text{photon}} - 1\rangle|N_{\text{phonon}} - 1\rangle$

[17] An example of this case is for the semiconductor Cu_2O.

if it proceeds by phonon absorption, or $|f\rangle|N_{\text{photon}}-1\rangle|N_{\text{ phonon}}+1\rangle$ it it proceeds by phonon emission. Here N_{photon} and N_{phonon} are the photon and phonon numbers. Because the net process must conserve energy, we must have $E_i+\hbar\omega_{\text{photon}}\pm\hbar\omega_{\text{phonon}}=E_f$, which implies $\hbar\omega_{\text{photon}}\pm\hbar\omega_{\text{phonon}}=E_g$. Note that this implies that the photon energy need not be the direct gap energies E_1 or E_2; the onset of absorption occurs for the smaller photon energies $\hbar\omega_{\text{photon}}=E_g\pm\hbar\omega_{\text{phonon}}$ as indicated in Fig. 27.13 (b).

Because the electron-photon matrix element $\langle f|W_{el-photon}|i\rangle=0$, the indirect allowed absorption coefficient is modeled quantum mechanically via higher order perturbation terms in Fermi's golden rule, as was discussed in Chapter 20, Equation 20.41. For this case, we use

$$\frac{1}{\tau(i\to f)}=\frac{2\pi}{\hbar}\Bigg|\underbrace{\langle f|W_{el-photon}+W_{el-phonon}|i\rangle}_{=0}+$$
$$\sum_m\frac{\langle f|W_{el-photon}+W_{el-phonon}|m\rangle\langle m|W_{el-photon}+W_{el-phonon}|i\rangle}{E_i-E_m}+...\Bigg|^2\times$$
$$\delta[E_c(\mathbf{k})-(E_v(\mathbf{k})+\hbar\omega_{\text{photon}}\pm\hbar\omega_{\text{phonon}}),\tag{27.55}$$

where $\langle f|W_{el-phonon}|i\rangle=0$ if $E_g\gg\hbar\omega_{\text{phonon}}$. Only the second order perturbation terms are retained as a sum over all intermediate states $|m\rangle$ that allow strict momentum conservation in every part of the step, and strict energy conservation between the initial and final states (for example $|m\rangle=|1\rangle,|2\rangle,...$). The cross terms of the type $\langle f|W_{el-photon}|m\rangle\langle m|W_{el-photon}|i\rangle$ and $\langle f|W_{el-phonon}|m\rangle\langle m|W_{el-phonon}|i\rangle$ are also zero. The term $\langle f|W_{el-photon}|m\rangle\langle m|W_{el-phonon}|i\rangle$ which represents a phonon absorption (or emission) followed by a photon absorption, and $\langle f|W_{el-phonon}|m\rangle\langle m|W_{el-photon}|i\rangle$ which represents the process of photon absorption followed by phonon absorption (or emission) are only non-zero terms. The absorption coefficient may then be evaluated by summing over both electron states $\mathbf{k}$ *and* phonon wavevectors β to yield the net 3D absorption coefficients

$$\alpha_0^c(\hbar\omega)=\frac{g_c q^2 m_v^{\frac{3}{2}} f_c\omega_{phonon}E_1(\hbar\omega\pm\hbar\omega_{phonon}-E_g)^2N_\pm}{32\pi\epsilon_0 n_r c m_e m_c^{\frac{1}{2}}\hbar\omega\lambda_c k_b T(E_1-\hbar\omega)^2},\text{ and}$$
$$\alpha_0^v(\hbar\omega)=\frac{g_c q^2 m_c^{\frac{3}{2}} f_v\omega_{phonon}E_2(\hbar\omega\pm\hbar\omega_{phonon}-E_g)^2N_\pm}{32\pi\epsilon_0 n_r c m_e m_v^{\frac{1}{2}}\hbar\omega\lambda_v k_b T(E_2-\hbar\omega)^2}.\tag{27.56}$$

Here g_c is the valley degeneracy of the final state $|f\rangle$, f_c, f_v are oscillator strengths, λ_c, and λ_v are the electron and hole mean free paths due to phonon scattering. Similar to Chapters 22 and 23, the squared electron-phonon interaction matrix elements are proportional to the phonon number $N_+=(e^{\hbar\omega_{\text{phonon}}/k_bT}-1)^{-1}$ for absorption, and $N_-=1+(e^{\hbar\omega_{\text{phonon}}/k_bT}-1)^{-1}$ for emission that includes the spontaneous and stimulated processes[18].

---

[18] Note that the indirect allowed absorption coefficient goes as the *square* of the excess photon energy. This is why experimental plots of the type in Fig. 27.13 (b) as a function of $\sqrt{\alpha(\hbar\omega)}$ show the intercepts at $E_g\pm\hbar\omega_{\text{phonon}}$, and can allow the measurement of the bandgap and the phonon energy simultaneously. See Exercise 27.9.

27.10 Exciton absorption

Though the absorption spectrum of silicon mostly follows Fig. 27.13 (b), that of germanium does not. The band-edge absorption spectra of direct-bandgap semiconductors such as GaAs also deviate from the calculated $\alpha_0(\hbar\omega) \propto \sqrt{\hbar\omega - E_g}$ dependence of Equation 27.40 at room temperature, and rather dramatically at low temperatures as indicated in Fig. 27.14 (b). The absorption coefficient is larger, and there are sharp peaks *below* the band-edge $\mathcal{M}_0$ critical point. These deviations of the band-edge absorption spectra is due due to the formation of **excitons**, a many-particle effect that is not captured in the preceding sections in this chapter.

Consider the absorption of a photon by a hydrogen atom. A photon of energy $\hbar\omega \geq R_y = 13.6$ eV ejects the electron from the atom by lifting it from the ground state energy to the continuum, ionizing the hydrogen atom. But several sharp absorption peaks appear well below this ionization energy, at $\hbar\omega = R_y/n^2$, where $n = 1, 2, \ldots$. These peaks result from changing the electron state from $E_1 \to E_2, \ldots$. These energies are "internal" excitations of the combined (electron + proton) system, which do not result in ionization. The energy is stored in the internal degrees of freedom of the electron, and may be released back by the emission of a photon.

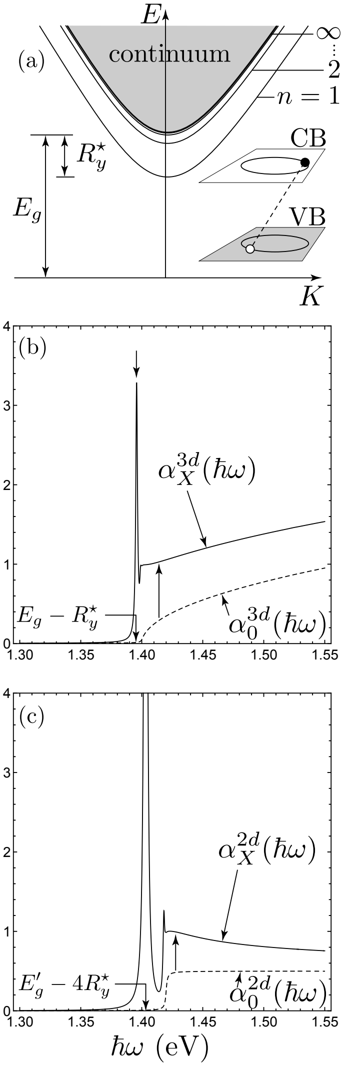


Fig. 27.14 (a) Exciton bandstructure and real-space representation as an electron-hole pair. (b) Absorption spectrum of direct-bandgap semiconductor with excitonic effect included for 3D carriers, and (c) the same for 2D carriers. A fundamental gap of $E_g = 1.4$ eV and exciton binding energy $R_y^\star = 4.5$ meV is assumed, and broadening is included.

The solid-state analogy of the above atomic process is the **exciton**, with the hole in the valence band of the semiconductor playing the role of the proton for the hydrogen atom, and the electron in the conduction band playing the role of the bound electron state of the hydrogen atom. To evaluate the absorption coefficient with exciton formation, it is necessary to retrace our steps, and now consider the **Coulomb attraction potential** $V(r) = -q^2/(4\pi\epsilon_s r)$ between the electron and the hole, that was neglected in the single-particle picture. In the single particle picture, the initial state was the valence band electron $\psi_{\mathbf{k}_v}(\mathbf{r}_v)$ of energy $E_v(\mathbf{k}_v)$ and the final state was the conduction band state $\psi_{\mathbf{k}_c}(\mathbf{r}_c)$ of energy $E_c(\mathbf{k}_c)$. The exciton state is a linear superposition of the conduction and valence band states: $\psi(\mathbf{r}_c, \mathbf{r}_v) = \sum_{\mathbf{k}_c, \mathbf{k}_v} A_{\mathbf{k}_c, \mathbf{k}_v} \psi_{\mathbf{k}_v}(\mathbf{r}_v) \psi_{\mathbf{k}_c}(\mathbf{r}_c)$ where $A_{\mathbf{k}_c, \mathbf{k}_v}$ are complex amplitudes. The combined energy of the exciton is $E(\mathbf{k}_c, \mathbf{k}_v)$. To find the allowed exciton energies, the following effective mass equation must be solved for E:

$$\Big(-\frac{\hbar^2}{2M}\nabla_R^2 - \frac{\hbar^2}{2m_r^\star}\nabla_r^2 + V(r)\Big)\psi(R, r) = (E - E_g)\psi(R, r), \tag{27.57}$$

where the transformation to the center of mass coordinates with $r = |\mathbf{r}_c - \mathbf{r}_v|$, $R = |\frac{m_c^\star \mathbf{r}_c + m_v^\star \mathbf{r}_c}{M}|$, $M = m_c^\star + m_v^\star$, and $m_r^\star = \frac{m_c^\star m_v^\star}{m_c^\star + m_v^\star}$ has been used. For the exciton, the combined motion of the electron-hole pair of mass M and coordinate R is factored out of the "internal" attraction $V(r) = -q^2/(4\pi\epsilon_s r)$ by writing $\psi(R, r) = \psi_R(R)\psi_r(r)$. Using[19] plane-waves $\psi_R(R) = \frac{e^{i\mathbf{K}\cdot\mathbf{R}}}{\sqrt{V}}$ for the combined motion of the exciton, the combined equation converts to the familiar hydrogen-atom form:

---

[19] The physics of excitons was first outlined theoretically by J. Frenkel in 1931, and by Wannier and Mott in 1937. Excitons were first observed in 1952 by Gross in Cu_2O.

$$\Big(-\frac{\hbar^2}{2m_r^\star}\nabla_r^2 - \frac{q^2}{4\pi\epsilon_s r}\Big)\psi_r(r) = (E-(E_g+\frac{\hbar^2K^2}{2M}))\psi_r(r),$$

$$\implies \boxed{E = E_g + \frac{\hbar^2K^2}{2M} - \frac{R_y^\star}{n^2},} \text{ and } \boxed{\psi(R,r) = \frac{e^{i\mathbf{K}\cdot\mathbf{R}}}{\sqrt{V}}\cdot\psi_{n,l,m}(r,\theta,\phi).} \tag{27.58}$$

The eigenvalues E are therefore composed of the bandgap E_g, the **continuous** exciton kinetic energy $\hbar^2K^2/2M$, and the **discrete** internal attraction energy $-R_y^\star/n^2$, in complete analogy to the hydrogen atom and the effective mass theory of donors as was discussed in Chapter 14. Fig. 27.14 (a) shows the resulting energy dispersion of the exciton energies. The eigenfunctions are composed of the corresponding plane-wave part and the hydrogenic bound part whose values at the origin $R=0$ are shown in Table 27.2. The value of the wavefunction at the origin is needed to calculate the absorption coefficient. The general formalism of calculation of the absorption coefficient is unchanged: as in the last section, it is still given by Equation 27.35, but with *different meanings* of the terms to account for excitons:

Table 27.2 Hydrogenic atom eigenfunctions and eigenvalues.

$a_B^\star$	$\frac{4\pi\hbar^2\epsilon_s}{q^2m_r^\star}$
$R_y^\star$	$\frac{m_r^\star q^4}{2\hbar^2(4\pi\epsilon_s)^2}$
3D	
Bound	
E_n	$-\frac{R_y^\star}{n^2}$
$\lvert\psi_n(0)\rvert^2$	$\frac{1}{\pi(a_B^\star)^3n^3}$
Continuum	
E	$\frac{\hbar^2k^2}{2m_r^\star}$
$\lvert\psi_E(0)\rvert^2$	$\frac{1}{R_y^\star 4\pi(a_B^\star)^3}[\frac{e^{\frac{\pi}{ka_0}}}{\sinh(\frac{\pi}{ka_0})}]$
2D	
Bound	
E_n	$-\frac{R_y^\star}{(n-\frac{1}{2})^2}$
$\lvert\psi_n(0)\rvert^2$	$\frac{1}{\pi(a_B^\star)^3(n-\frac{1}{2})^2}$
Continuum	
E	$\frac{\hbar^2k^2}{2m_r^\star}$
$\lvert\psi_E(0)\rvert^2$	$\frac{1}{R_y^\star 4\pi(a_B^\star)^3}[\frac{e^{\frac{\pi}{ka_0}}}{\cosh(\frac{\pi}{ka_0})}]$

$$\alpha(\hbar\omega) = \frac{\pi q^2}{\eta c\epsilon_0 m_e^2\omega}\cdot\frac{2}{V}\sum_{\mathbf{k}}|\langle f|\hat{\mathbf{p}}|i\rangle|^2\delta[E_f-(E_i+\hbar\omega)] \rightarrow$$

$$\boxed{\alpha_X(\hbar\omega) = \frac{\pi q^2}{\eta c\epsilon_0 m_e^2\omega}\cdot\frac{2}{V}\sum_{\mathbf{k}}|\psi_n(0)|^2\delta[(E_g+\frac{\hbar^2K^2}{2M}-\frac{R_y^\star}{n^2})-(0+\hbar\omega)].} \tag{27.59}$$

Here the initial state is of energy $E_i=0$ with all electrons in the valence band and none in the conduction band. The final state is that of the exciton of energy from Equation 27.58. Exercise 27.10 shows how the squared momentum matrix element term $|\langle f|\hat{\mathbf{p}}|i\rangle|^2 \rightarrow |\psi_n(0)|^2$ converts to the value of the squared wavefunction at the origin. Using both the discrete and continuum states listed in Table 27.2 in boxed Equation 27.59, the absorption coefficient due to excitons in 3D is[20]

$$\alpha_X^{3d}(\hbar\omega) = \frac{A_0}{\pi(a_B^\star)^3R_y^\star}\Big(\sum_{n=1}^{\infty}\frac{2}{n^3}\frac{\gamma/\pi}{(u+\frac{1}{n^2})^2+\gamma^2}+\frac{1}{2\pi}\int_0^\infty\frac{du'}{\pi}\frac{\gamma\sqrt{u'}\mathcal{S}_{3d}(u')}{(u'-u)^2+\gamma^2}\Big)$$

$$\text{where } u = \frac{\hbar\omega-E_g}{R_y^\star}, \mathcal{S}_{3d}(u) = \frac{2\pi/\sqrt{u}}{1-e^{-2\pi/\sqrt{u}}}, \text{ and } A_0 = \frac{\pi q^2|\hat{e}\cdot\mathbf{p}_{cv}|^2}{\eta c\epsilon_0\omega m_e^2}. \tag{27.60}$$

Here, a broadening parameter γ has been used. Fig. 27.14 (b) shows this exciton absorption spectrum in comparison to the single-particle absorption spectra $\alpha_0^{3d}(\hbar\omega) = \frac{A_0}{\pi(a_B^\star)^3R_y^\star}\cdot\frac{1}{2\pi}\sqrt{u}$. The continuum absorption spectrum is increased by a factor of 2π and the sharp $\mathcal{M}_0$

---

[20] The term $\mathcal{S}_{3d}(u)$ is the Sommerfeld enhancement factor in the absorption coefficient due to the continuum exciton states. Excitonic effects vanish for $\mathcal{S}\rightarrow 1$, and if the bound states are neglected.

point is removed by this many-body excitonic peak. The experimental absorption spectrum of bulk GaAs indeed looks like the solid excitonic spectrum of Fig. 27.14 (b) rather than the dashed single-particle spectrum, emphasizing the importance of excitons in the light-matter interaction.

For excitons confined in 2D in quantum wells, the problem maps to a 2D hydrogen atom analog; Table 27.2 lists the exciton energies and functions, which lead similarly to the absorption coefficient

$$\alpha_X^{2d}(\hbar\omega) = \frac{A_0}{\pi(a_B^\star)^2 R_y^\star}\Big(\sum_{n=1}^{\infty}\frac{2}{(n-\frac{1}{2})^3}\frac{\gamma/\pi}{(u+\frac{1}{(n-\frac{1}{2})^2})^2+\gamma^2} + \frac{1}{2}\int_0^{\infty}\frac{du'}{\pi}\frac{\gamma\mathcal{S}_{2d}(u')}{(u'-u)^2+\gamma^2}\Big) \text{ where } \mathcal{S}_{2d}(u) = \frac{2}{1+e^{-2\pi/\sqrt{u}}}, \tag{27.61}$$

with a modified Sommerfeld enhancement factor. The 4× enhancement of the binding energy over the 3D value leads to stronger sub-E_g absorption, as indicated in Fig. 27.14 (c).

27.11 Franz–Keldysh effect

The absorption coefficient in bulk semiconductors and in lower dimensional heterostructures is modified by the application of large electric or magnetic fields. The modification of the semiconductor absorption coefficient and the related optical properties by electric field forms the physical basis of several electro-optical devices such as modulators and detectors that are used in photonic communications. Examples of such devices are discussed in Chapter 28. Here we discuss the effect of the electric field on the optical properties quantitatively.

In Chapter 25, we encountered the Shubnikov de Haas effect, in which the electrical conductivity in a semiconductor became oscillatory with a varying magnetic field. This quantum oscillation occurs because the magnetic field clumps the DOS into Landau levels, and because the electronic conductivity $\sigma \propto g(E_F) \propto \cos(\frac{2\pi E_F}{\hbar\omega_c})$ is proportional to the DOS at the Fermi level, the changing magnetic field sweeps Landau level peaks in the DOS across the Fermi level, causing the conductivity to oscillate. An optical analogy of the Shubnikov de Haas effect occurs in semiconductors in the presence of a large *electric* field. This is called the **Franz–Keldysh effect**: the conductivity (or transmission) of photons through the semiconductor of energy larger than the bandgap oscillates with the photon energy for a fixed electric field (Fig. 27.15 (a)), or with electric field for a fixed photon energy (Figures 27.15 (c)). More importantly, photons of energy $\hbar\omega \lesssim E_g$ less than but close to the bandgap can now be absorbed (Figures 27.15 (b)). The **sub-bandgap absorption coefficient depends exponentially on the electric field magnitude** F as $\alpha(F) \propto \exp(-F/F_0)$, where $F_0 = \frac{4}{3}\sqrt{\frac{2m_r^\star}{\hbar^2}}\frac{(E_g-\hbar\omega)^{3/2}}{q}$ is a characteristic electric field for such

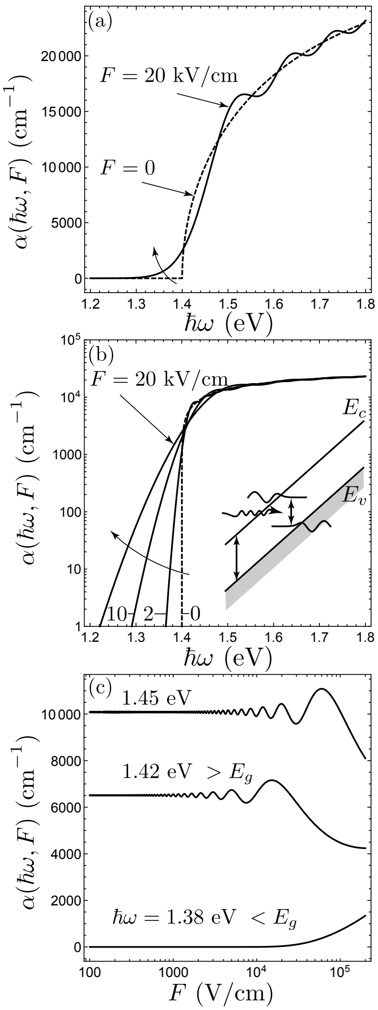


Fig. 27.15 (a) Modification of the 3D single-particle absorption spectrum by an electric field. (b) Same as (a) in log scale highlighting the large change in sub-bandgap absorption, which is a photon-assisted interband tunneling process. (c) Variation of absorption coefficient as a function of the electric field at three photon energies. For photon energies below the zero-field edge it is a monotonic increase with field, and for above zero-field edge it is oscillatory with field. Parameters used for the plots are $E_g = 1.4$ eV, $m_c^\star = 0.067m_e$, $m_v^\star = 0.53m_e$, $\eta = 3.3$, and $2|p_{cv}|^2/m_e = 20$ eV.

absorption. This dependence on the electric field is reminiscent of the interband WKB tunneling probability of Chapter 24, and for good reason: here the electron is effectively tunneling from the valence band to the conduction band, but now aided by a photon. The detailed derivation proceeds similarly to the exciton case by considering the electron-hole pair, but now the interaction term is due to the *externally* applied electric field along the z-axis, leaving the (x, y) plane free:

$$\Big(-\frac{\hbar^2}{2m_r^\star}\frac{\partial^2}{\partial z^2}+qFz\Big)\psi(z)=\underbrace{[E-\frac{\hbar^2}{2m_r^\star}(k_x^2+k_y^2)]}_{E_z}\psi(z), \tag{27.62}$$

which, with the substitutions $l_F = (\frac{\hbar^2}{2m_r^\star qF})^{\frac{1}{3}}$ and $\mathcal{Z} = (z - \frac{E_z}{qF})/l_F$ transforms to the equation $\partial^2\psi(\mathcal{Z})/\partial\mathcal{Z}^2 = \mathcal{Z}\psi(\mathcal{Z})$. The solutions to this differential equation are the well-known Airy functions, which are oscillatory for $\mathcal{Z} < 0$ and exponentially decaying for $\mathcal{Z} > 0$. Using a similar strategy as Equation 27.60 for excitons, by summing over all (k_x, k_y) states perpendicular to the field and the continuum of states of energy E_z, the absorption coefficient is obtained as

$$\alpha(\hbar\omega, F) = \frac{A_0}{2\pi}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{\hbar\omega_F}[\mathrm{Ai}'^2[\epsilon] - \epsilon\mathrm{Ai}^2(\epsilon)]$$
$$\text{where } \hbar\omega_F = (\frac{\hbar^2q^2F^2}{2m_r^\star})^{\frac{1}{3}}, \text{ and } \epsilon = \frac{E_g - \hbar\omega}{\hbar\omega_F}, \tag{27.63}$$

where Ai$'$ is the derivative of the Airy function[21]. Fig. 27.15 shows plots of this modification of the absorption coefficient with electric field. Note the strong increase in the sub-bandgap absorption with electric field, and the oscillatory dependence of the absorption coefficient with electric field. Because the reflectivity is related to the absorption coefficient via the Kramers–Kronig relations, Franz–Keldysh oscillations are also observed in the refectivity (see Fig. 27.16).

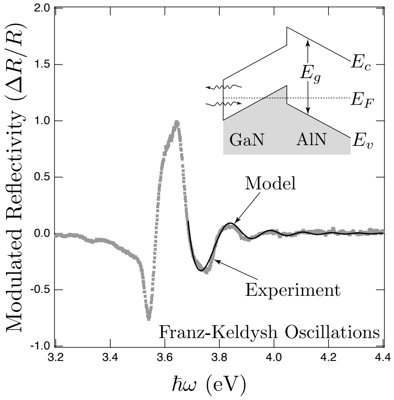


Fig. 27.16 Franz–Keldysh oscillations observed in contactless electroreflectance from a thin GaN layer grown on top of AlN. A large polarization-induced electric field exists in the GaN layer as indicated in a schematic energy band diagram in the inset. Data provided by Ł. Janicki, R. Chaudhuri and R. Kudrawiec (Phys. Stat. Solidi RRL, 2021 2000573).

The field strengths to achieve appreciable Franz–Keldysh effect in 3D semiconductors may be large, sometimes close to the breakdown fields of the semiconductors. On the other hand, the effect is rather strong and achievable in 2D quantum wells, where it goes by the name of the **quantum-confined Stark effect, or QCSE**[22].

Absorption properties in a large magnetic field: We saw in Chapter 25 that at high magnetic fields, the allowed eigenvalues in a semiconductor band change to $E(k_x, k_y, k_z) \rightarrow E(k_z) + (n + \frac{1}{2})\hbar\omega_c$ where $\omega_c = qB/m^\star$ is the cyclotron frequency. The density of states clumps up into Landau levels. Optical transitions can occur either between the Landau levels of the same band, which are spaced by $\hbar\omega_c$, or from those in one band to another, which are spaced by $E_g + \hbar g_c \frac{qB}{m_c^\star} + \hbar g_v \frac{qB}{m_v^\star}$ where g_c and g_v are the magnetic g-factors of the bands. Magneto-optical transitions of this kind were used, for example, in the earliest

[21] Note the similarity of this result to the Kubo formalism of interband tunneling discussed in Chapter 24, Section 24.13.

[22] Both Franz–Keldysh and QCSE effects are affected strongly by exciton effects in 3D and 2D. The electric field and the Coulomb attraction of electrons and holes should then be treated on equal footing. We refrain from presenting the full-blown treatment here.

cyclotron resonance measurements to deduce the bandstructures of germanium and silicon.

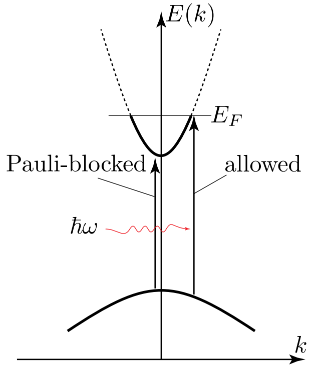


Fig. 27.17 The Burstein–Moss blue shift in the minimum $\hbar\omega$ for optical transition by Pauli-locking due to band-filling.

Burstein–Moss shift: Fig. 27.17 shows that the absorption of photons near and slightly above the bandgap is forbidden if the Fermi level is high in the conduction band. The factor $f_v(\mathbf{k}) - f_c(\mathbf{k})$ in the expression for the absorption coefficient ensures that because of the Pauli exclusion principle, it is impossible to put another electron in the state that is already occupied. Though in later chapters we will see that this difference function of occupation probabilities may be dynamically changed in non-equilibrium by carrier injection or extraction, such blocking is indeed seen in heavily doped semiconductors at equilibrium. The minimum photon energy that can be absorbed becomes $\hbar\omega \geq E_g + \frac{\hbar^2 k_F^2}{2m_r^\star}$, where $\frac{\hbar k_F^2}{2m_c^\star} = E_F - E_c$, and $k_F = (3\pi^2 n)^{1/3}$, where n the density of free carriers in the parabolic conduction band. A similar shift is observed also for heavily p-type doped bulk semiconductors. The blue shifts can be substantial in narrow gap semiconductors with light effective masses; for high carrier densities, both band nonparabolicity, and bandgap shrinkage due to heavy doping must be accounted for. Note that the change can also be dynamic, but the injection of electrons or holes, which makes it a powerful method to make the semiconductor opaque or transparent to desired wavelengths.

27.12 Intersubband absorption

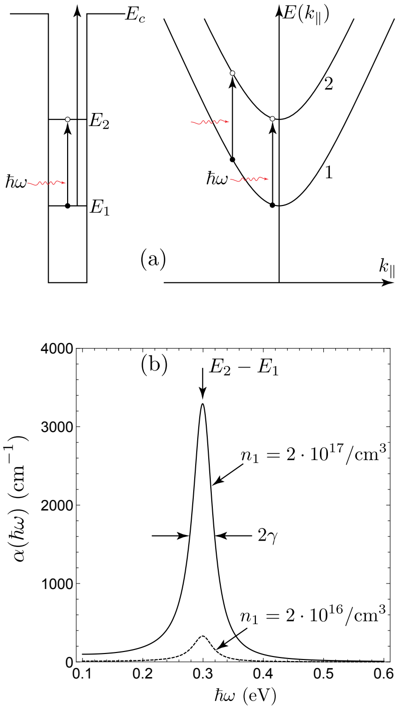


Fig. 27.18 (a) Intersubband absorption of photons in a quantum well, and (b) the resulting absorption coefficient calculated for $E_2 = 0.4$ eV, $E_1 = 0.1$ eV, $z_{if} = 2$ nm, $\eta = 3.3$, $\gamma = 20$ meV. The volume carrier density n_1 of the lower subband is indicated and the upper subband density $n_2 \approx n_1/100$ is assumed.

In prior sections we have discussed photon absorption by electron transition from the valence to the conduction band, which are interband processes. Fig. 27.18 (a) indicates that absorption is also possible by transition of electrons between different subbands of the same band, and Fig. 27.18 (b) shows an example of the resulting absorption spectra. The spectrum is peaked for photons that match the intersubband transition energy $\hbar\omega = E_2 - E_1$ and is more atomic-like than band-like. The strength of absorption is strongly dependent on the difference of occupation $n_1 - n_2$ of the two subbands. The intersubband absorption coefficient can be reasonably strong if the subbands are equidistant in energy over a large portion of the k-space.

To compute the intersubband absorption coefficient, we identify $E_i = E_1 + \frac{\hbar^2 k_\parallel^2}{2m_c^\star}$ and $E_f = E_2 + \frac{\hbar^2 (k'_\parallel)^2}{2m_c^\star}$ as the dispersions of the two lowest subbands of the quantum well, and $\psi_i = \frac{e^{i\mathbf{k}_\parallel \cdot \mathbf{r}_\parallel}}{\sqrt{A}} \cdot \chi_1(z) \cdot u_c(\mathbf{r})$ and $\psi_f = \frac{e^{i\mathbf{k}'_\parallel \cdot \mathbf{r}_\parallel}}{\sqrt{A}} \cdot \chi_2(z) \cdot u_c(\mathbf{r})$ are the corresponding wavefunctions written explicitly as the product of the envelope and Bloch lattice components. We use the equivalent dipole form of the momentum matrix element shown in Equation 27.9 in the form $\mathbf{p}_{21} = im_e\omega_{21}\langle 2|\mathbf{r}|1\rangle$ to get $|\mathbf{p}_{21}|^2 = m_e^2\omega_{21}^2 |z_{21}|^2 \delta_{\mathbf{k}_\parallel, \mathbf{k}'_\parallel}$, where $\hbar\omega_{21} = E_2 - E_1$, $z_{21} = \int dz \chi_2(z) z \chi_1(z)$ is the dipole length (which evaluates to a fraction of the quantum well width[23]), and the Kronecker delta allows only vertical transitions to conserve momentum. Assuming f_1 and f_2 to be the occupation functions of states in subband 1 and 2 respectively, and using the general form of the absorption coefficient we obtain

$$\begin{aligned}\alpha(\hbar\omega) &= \frac{\pi q^2}{\eta c\epsilon_0 m_e^2\omega}\frac{2}{V}\sum|\langle f|\mathbf{p}|i\rangle|^2\delta[E_f-(E_i+\hbar\omega)]\times(f(E_i)-f(E_f))\\ &= \frac{\pi q^2|z_{21}|^2\omega_{21}^2}{\eta c\epsilon_0\omega}\frac{2}{V}\sum\underbrace{\delta[(E_2-E_1)-\hbar\omega)]}_{\frac{\gamma/\pi}{(E_2-E_1-\hbar\omega)^2+\gamma^2}}\times(f(E_i)-f(E_f))\\ &= \frac{\pi q^2|z_{21}|^2\omega_{21}^2}{\eta c\epsilon_0\omega}\frac{\gamma/\pi}{(E_2-E_1-\hbar\omega)^2+\gamma^2}\underbrace{\frac{2}{V}\sum(f(E_i)-f(E_f))}_{n_1-n_2}\\ &\implies \boxed{\alpha(\hbar\omega)=\frac{q^2|z_{21}|^2\omega_{21}^2}{\eta c\epsilon_0\omega}\frac{\gamma}{(E_2-E_1-\hbar\omega)^2+\gamma^2}(n_1-n_2),}\end{aligned} \tag{27.64}$$

where γ is fundamentally the inverse of the upper state lifetime (decoherence rate). The broadening of the energy levels occurs due to interface roughness. In multiple quantum well cases there is additional broadening due to inter-well coupling of the higher energy states that form broader minibands. The volume density difference n_1-n_2 implies that there is no intersubband transition if both subbands are empty, of if both are filled for the same $k_{\|}$. Fig. 27.18 shows a plot of the intersubband absorption spectrum for parameters indicated in the caption. The linewidth of the transition is 2γ and the peak value is $\alpha_{max}=\frac{q^2|z_{21}|^2\omega_{21}}{\eta c\epsilon_0\gamma}(n_1-n_2)$, which highlights the sharp increase possible by decreasing γ.

Because intersubband energies are controlled by the quantum well thicknesses instead of the bandgap, several semiconductor heterostructure combinations with varying properties can be used for transitions at roughly the same energy. The energies for intersubband transitions typically fall in the IR and long-wavelength regime of the spectrum, and have applications such for detection of atmospheric pollutants[24].

27.13 Free carrier, and impurity absorption

The mechanism of absorption of photons in a semiconductor is not limited to the interband and intersubband transitions discussed in the previous sections. The generalized semiconductor absorption spectra of Fig. 27.4 indicates a "floor" that goes approximately as $\alpha(\omega)\propto 1/(\hbar\omega)$. This is primarily due to **free-carrier absorption** processes. The figure also indicates **impurity absorption** peaks due to transitions between dopant → band states, or dopant → dopant states. In this section, we discuss these optical processes, and touch upon their im-

[23] Note that $\mathbf{r}=x\hat{x}+y\hat{y}+z\hat{z}$, and because of plane wave states in the (x,y) plane, the x and y dipole matrix elements are zero and only the z-component survives. This immediately means that the electric field component of the photon must have a non-zero component in the z-direction to cause an intersubband transition. Speaking classically, to move an electron in the z-direction requires a force in that direction. Furthermore, the normalized cell-periodic parts $u_c(\mathbf{r})$ integrate to 1 over the crystal.

[24] Photodetector applications are discussed in Chapter 28. Note that when $n_1<n_2$, the absorption coefficient $\alpha(\hbar\omega)$ becomes negative: this should not cause alarm and is perfectly OK (and very interesting!). A negative absorption coefficient implies optical gain, which leads to intersubband quantum cascade lasers that are discussed in Chapter 29.

portance in photonic devices.

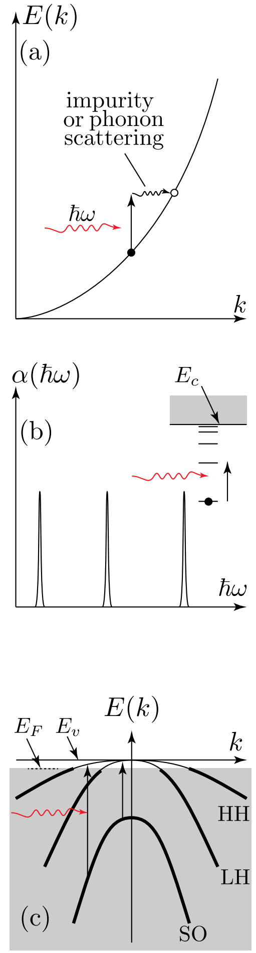


Fig. 27.19 (a) Free carrier absorption requires a scattering event for momentum conservation. (b) Absorption by an electron on a donor site exhibits a hydrogen-atom like spectrum with effective band and crystalline properties. (c) Interband absorption between the valence bands in a heavily p-doped semiconductor.

Free carrier absorption: The free carrier absorption process is indicated in Fig. 27.19 (a). It is clear that because of the restriction of vertical transitions, free carriers in the band cannot absorb a photon without the help from either phonons, or impurity scattering that supplies the momentum needed to access another allowed electron state at a higher energy. The method for the evaluation of the absorption coefficient mirrors the classical Drude–Lorentz oscillator approach that was used in Chapter 26. The quantum mechanical approach also yields, using $\omega_p^2 = nq^2/\epsilon_0 m_e$ as the plasma frequency of the free carriers, and τ the momentum scattering time the absorption coefficient,

$$\alpha(\hbar\omega) = \frac{1}{\eta c \tau} \cdot \frac{\frac{q^2 n}{\epsilon_0 m^\star}}{\omega^2 + \frac{1}{\tau^2}}, \tag{27.65}$$

which decreases as the square of the photon energy for $\omega\tau >> 1$. This condition is satisfied for for IR-UV light for semiconductors with moderate to high mobilities. As an example, for 1.0 eV photons, the absorption by $n = 10^{18}/\text{cm}^3$ carriers in a semiconductor of effective mass $m^\star = 0.1 m_e$ and scattering time $\tau = 10^{-13}$ s leads to an absorption coefficient of $\alpha \sim 1.4$ cm^{-1}, but for a ten times smaller photon energy of 100 meV, the coefficient increases by 100 times to ~ 140 cm^{-1}.

Impurity absorption: Fig. 27.19 (b) shows the effective hydrogenic states of a donor impurity below the conduction band edge. Because this is the solid-state analogy of the hydrogen atom with re-scaled Rydberg energies, sharp absorption lines are observed for the transitions from filled to empty bound states, or to the continuum states in the band. The absorption coefficient has the exact same form of the hydrogen atom, but with broadening associated with lattice vibrations, or the effect of nearby impurities. Such transitions were explored in the past, and are making a comeback for potential applications in semiconductor qubits. Since the donor ionization energies are in the few to 10s of meV, these qubits need to be operated at low temperatures.

Inter-valence band absorption: Fig. 27.19 (c) shows another kind of absorption that is possible between the three valence bands in a degenerately doped p-type semiconductor in which the Fermi level is inside in the valence band. Because most states above the Fermi level are free, photons that match the energy difference between the branches of the valence bands can now be absorbed if the angular momentum selection rules allow the transition. Note that these transitions are vertical in **k** unlike those of Fig. 27.19 (a).

27.14 Photoelectron spectroscopy

Photons of much higher energies than the bandgap of the semiconductor can eject electrons from the valence bands into the continuum. The electrons ejected from the surface of the semiconductor carry information about their momentum $\mathbf{k}$: by measuring the energies and momenta of these ejected electrons, the bandstructure $E(\mathbf{k})$ of the semiconductor itself can be measured. This is the principle behind the Angle-Resolved Photoelectron Spectroscopy (or ARPES) technique, which has helped understand the bandstructures of various classes of crystalline materials ranging from metals to semiconductors and insulators. This is really a much more refined form of the photoelectric effect itself, for which Einstein provided the first theoretical explanation and put quantum mechanics on firm ground, as was discussed in Chapter 3.

In Fig. 13.13 in Chapter 13, we encountered an example of the valence bands of GaN measured by ARPES alongside the calculated bandstructure. The agreement highlights the accuracy enabled by significant advances in the experimental techniques, and also in the theoretical modeling of the bandstructure. The combined light-matter interaction study therefore is gradually taking semiconductors towards the goal of a comprehensive understanding of a crystal at par with that of the hydrogen atom, whose every eigenvalue and eigenfunction is known nearly exactly, and there is theory that supports every fine structure and splitting. This quest continues for semiconductor physics through the combined theoretical and experimental advances.

27.15 Chapter summary section

In this chapter, we learned:

- The quantum mechanical recipes for treating light-matter interaction in any solid via the light-matter perturbation term $\frac{q}{m_e}\mathbf{A}\cdot\mathbf{p}$.
- The importance and the evaluation of the absorption coefficient of semiconductors.
- The principle of vertical transitions due to photon absorption, and why direct bandgap semiconductors have much larger band-edge absorption coefficients at band-edge energies than indirect-bandgap semiconductors.
- Indirect bandgap semiconductors can still absorb photons of energies close to the bandgap by the assistance of phonon absorption or emission.

- The bandstructure over the entire Brillouin zone allows for several critical points where the bands become parallel and there are sharp Van Hove singularities in the absorption spectra. The shape of the absorption and emission spectra near the critical points provide some of the most accurate tests for bandstructure calculations.
- In addition to single-particle interband transitions, many-particle effects such as exciton formation significantly modify the absorption spectra of semiconductors.
- The absorption spectrum of semiconductors can be changed dynamically by electric fields, by magnetic fields, or by carrier injection.
- Intersubband transitions, and also transitions due to impurities and free carriers play important roles, in addition to interband transitions in semiconductor photonic devices.

Further reading

Fundamentals of Semiconductors by P. Yu and M. Cardona has excellent chapters on optical properties of semiconductors that covers more ground than other textbooks. **Optical Properties of Solids** by M. Fox is a very readable and comprehensive textbook on light-matter interactions, which describes a wealth of experimental methods. **Optical Processes in Semiconductors** by J. Pankove is a classic text that remains modern in its presentation. **Physics of Photonic Devices** by S. L. Chuang provides a rigorous treatment of the photonic properties of semiconductors from the point of view of useful optical devices.

Exercises

(27.1) **Hydrogen atom optical transitions**

In this problem, we investigate an optical transition in the hydrogen atom, and discuss how the emission of photons from a single atom is different from the emission from a semiconductor crystal. Consider a hydrogen atom in which at $t = 0$ the electron is in the $|2p\rangle$ state with $\psi_{2p}(r,\theta,\phi) = \frac{1}{\sqrt{32\pi a_B^5}} r e^{-r/2a_B} \cdot \cos\theta$. For time $t > 0$ the electron can make a transition to the $|1s\rangle$ ground state with $\psi_{1s}(r,\theta,\phi) = \frac{1}{\sqrt{\pi a_B^3}} e^{-r/a_B}$, where $E_{2p} - E_{1s} = \frac{3q^2}{32\pi\epsilon_0 a_B}$ by emitting a photon of energy $\hbar\omega = E_{2p} - E_{1s}$.

(a) Using the relation between the optical dipole matrix element and the momentum matrix element, show that the dipole along $\hat{e}$ direction is

$$\int d^3r \psi_{1s}^{\star}(-q\mathbf{r})\psi_{2s} = -\frac{128\sqrt{2}}{243} q a_B \hat{e}. \qquad (27.66)$$

(b) By summing over all final photon states in Fermi's golden rule, show that the transition rate is

$$\frac{1}{\tau(2s \to 1s)} = (\frac{2}{3})^8 (\frac{q^2}{4\pi\epsilon_0 \hbar c})^3 \frac{m_e q^4}{(4\pi\epsilon_0)^2 \hbar^3}. \quad (27.67)$$

(c) Show that the lifetime is 1.6 ns, and the corresponding linewidth of the transition is 0.41 μeV.

(d) Argue why an atom can emit only one photon at a time, and how is this property different from the case in a solid.

(27.2) **Lorentzian lineshapes, homogeneous and inhomogeneous broadening**

(a) Consider the case where instead of one hydrogen atom, there are a large number in state $|2p\rangle$ at time $t = 0$. If all atoms have equal opportunity to relax to the $|1s\rangle$ state by emitting photons of energy $\hbar\omega = E_{2p} - E_{1s}$, then the resulting emission spectrum is still centered around $\hbar\omega = E_{2p} - E_{1s}$, but is broadened compared to that of a single atom. Show, using arguments similar to what was used in Chapter 25, Equation 25.36, that the spectrum has a Lorentzian lineshape. This is an example of **homogeneous broadening**.

(b) Now consider the case when there are several species of atoms, or if the atoms are colliding with each other or moving at various velocities. Explain why this causes several center frequencies of emission, and therefore leads to a broadened emission spectrum. This is called **inhomogeneous broadening**.

(27.3) **Optical property evaluation using the pseudopotential bandstructure**

This problem may be treated as a project, since it requires a longer computational effort than other problems. Refer to Chapter 13 and evaluate the pseudopotential bandstructures of silicon, GaAs, and GaN.

(a) Plot the energy bandstructures of each similar to those shown in Chapter 16.

(b) Plot the squared wavefunctions of states at the Γ-point of the conduction band and the valence band for each semiconductor. This is a 3D density, so make density plots in 3D, and also plot the contour plots in the 100 and 111 planes for silicon and GaAs, and the 0001 plane and 10-10 planes for GaN.

(c) Compute the optical matrix element $\mathbf{p}_{cv}$ using the evaluated wavefunctions for the Gamma-point, and points near it.

(d) Proceed to evaluate the optical absorption coefficient for each of the three semiconductors. For silicon, see how close your absorption spectrum looks to what is shown in Fig. 27.3 (b).

(27.4) **Inter- and intra-conduction band absorption in β-Ga_2O_3**

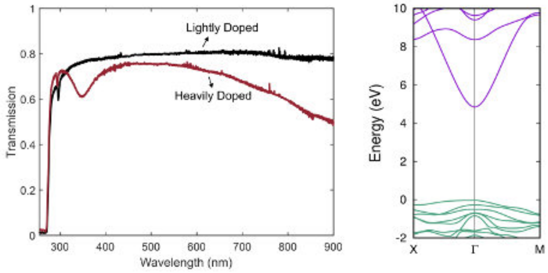


Fig. 27.20 Left: Experimentally measured room temperature optical transmission spectra for a heavily doped and a lightly doped β-Ga_2O_3 semiconductor crystal for light propagating along the [102] axis. Right: Calculated bandstructure of β-Ga_2O_3. Data kindly shared by Arjan Singh of Cornell University, adapted from Applied Physics Letters, volume 117, page 072103 (2020). Bandstructure calculated by Dr. Hartwin Peelaers of the University of Kansas.

Fig. 27.20 (left) shows the optical transmission spectra measured for the ultra wide-bandgap semiconductor β-Ga_2O_3 for light propagating along the [102] axis, and its calculated energy bandstructure is on the right. Answer the following questions:

(a) Is β-Ga_2O_3 a direct-bandgap or indirect-bandgap semiconductor?

(b) Identify the smallest energy valence to conduction band transition at the Γ-point of the bandstructure with an arrow, and verify the measured absorption edge wavelength indeed corresponds to this energy.

(c) In the absorption spectrum around 300 nm, a strong absorption dip is seen in heavily doped β-Ga_2O_3, which goes away for a lightly doped sample. This transition is identified to be due to inter-conduction band transitions. Identify this transition in the bandstructure with an arrow, verify its energy, and explain why this absorption dip is not observed in the lightly doped sample.

(d) The heavily doped sample shows a strong absorption at longer wavelengths. This increase in absorption is due to intra-conduction band free-carrier absorption. Sketch in the bandstructure which transitions explain this process, and explain again why it vanishes for the lightly doped sample.

The experimental data and the problem formulation was kindly shared by Arjan Singh of Cornell University.

(27.5) Optical absorption in monolayer graphene

The absorption coefficient of a monolayer of graphene is famously the simple relation $\alpha_0(h\nu) = \pi\alpha$, where $\alpha = q^2/(4\pi\epsilon\hbar c)$ is the fine structure constant, nearly ”independent” of the frequency (or wavelength) of light. The fine structure constant has the same symbol as the absorption coefficient, do not be confused!

(a) Using the linear dispersion bandstructure of graphene, prove following the same line of derivation of parabolic bandstructures why this must be the case.

(b) Note that unlike the 3D semiconductor case where $\alpha_0(h\nu)$ has units of (length)$^{-1}$, for graphene it is dimensionless. Explain why it is so, and estimate the numerical value of the absorption coefficient for layer of graphene in vacuum. It is remarkable that a single atomic monolayer can absorb that much of photons!

(c) Since a layer of graphene may be sandwiched between two dielectric regions, the fine-structure ”constant” is not necessarily a constant, but can be engineered. Explain what this implies for absorption by a single layer of graphene.

(d) Under what conditions will the simple relation $\alpha_0(h\nu) = \pi\alpha$ fail? The answer may be found in the full bandstructure of graphene.

(27.6) Photoluminescence, lattice and electron temperatures

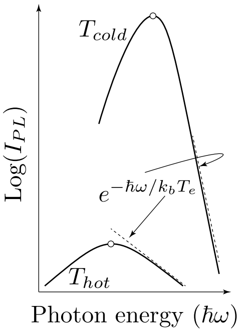


Fig. 27.21 Schematic representation of the photoluminescence spectra with the intensity in the log scale shown at two temperatures.

Fig. 27.21 schematically shows how the photoluminescence spectra for semiconductors are used to extract the effective temperature of electrons and holes in comparison to the temperature of the crystal lattice. Answer the following:

(a) Since luminescence is the inverse process of absorption, write an expression in terms of the equilibrium absorption coefficient $\alpha_0(\hbar\omega)$ and the occupation functions of the conduction and valence bands f_c, f_v for the emission spectrum. Show that the (stimulated) emission spectrum is given by $I_{PL}(\hbar\omega) \approx \alpha_0(\hbar\omega) f_c(1 - f_v)$. In real semiconductors, there is a sub-bandgap broadening, which we neglect in this problem.

(b) Verify that low-level excitation when the quasi-Fermi levels of electrons and holes stay within the bandgap of the semiconductor, the PL spectra is approximately expressed as $I_{PL} \approx \alpha_0(\hbar\omega) e^{-\hbar\omega/k_b T_e}$, where T_e is the temperature characterizing the energy of electrons in the conduction band.

(c) The bandgap of the semiconductor changes with the lattice temperature T_L according to the Varshni relation $E_g(T_L) = E_b(0) - \frac{\alpha T_L^2}{T_L + \beta}$, where $E_g(0)$ is the bandgap at $T_L = 0$ K, and α and β are material properties. Therefore, the peak of the PL spectra shifts with temperature which is measured in the experiment. Show how the PL peak will shift as a function of temperature.

(d) Since the high-energy tail is proportional to T_e, an Arrhenius fit to the high-energy PL tail gives a direct measure of the electron temperature. Discuss how this technique may be used to measure hot-carrier temperatures, when electrons are not in equilibrium with the lattice.

(27.7) **Van Hove singularities**

When two bands between which optical transitions are allowed become parallel in $E(k)$ vs. k, a large number of optical transitions are possible. This The general principle leads to Van Hove singularities in the optical absorption spectra. The small singularity regions in the Brillouin zone dominate the absorption spectra in magnitude. Explore the rich varieties of Van Hove singularities based on the exact nature of the bandstructures of semiconductors discussed in this book. Pay particular attention to warped bandstructures in which these regions form connected lines or surfaces in the Brillouin zones, and identify the crystals in which they occur. Explain why the observation of such singularities in the optical absorption spectra provides a stringent verification of theoretical bandstructure models of semiconductors and other crystals over a large energy range, far from the band-edges.

(27.8) **Forbidden direct and indirect transitions**

Equation 27.54 gives the absorption coefficient for a forbidden direct transition, which is understood by answering the following questions:

(a) In Equation 27.28 it was found that the optical matrix element split into two terms: except for the slowly varying $e^{i\mathbf{k}\cdot\mathbf{r}}$ term, the first term was the integral of the direct product of the valence and conduction band Bloch functions $\int d^3r u_c^\star(\mathbf{r}) u_v(\mathbf{r})$, and the second with the momentum operator nestled between them: $\int d^3r u_c^\star(\mathbf{r}) \hat{\mathbf{p}} u_v(\mathbf{r})$. Argue why if the valence band edge states are composed completely of $|p\rangle$ orbitals and the conduction band edge states completely of $|s\rangle$ orbitals, then $\int d^3r u_c^\star(\mathbf{r}) u_v(\mathbf{r})$ is indeed zero. But if the states are linear combinations of the type $c_s|s\rangle + c_p|p\rangle$, $\int d^3r u_c^\star(\mathbf{r}) u_v(\mathbf{r})$ is non-zero.

(b) On the other hand if both the valence band and conduction band states are of $|s\rangle$ or $|p\rangle$ orbital symmetry, then argue why the "direct" term $\int d^3r u_c^\star(\mathbf{r}) \hat{\mathbf{p}} u_v(\mathbf{r})$ vanishes but the "forbidden" term $\int d^3r u_c^\star(\mathbf{r}) u_v(\mathbf{r})$ is large. Which semiconductors show this behavior?

(c) Show that if we assume $f' = |\int d^3r u_c^\star(\mathbf{r}) u_v(\mathbf{r})|^2 \neq 0$ because of the orbital mixtures that allow this to be the case, then the absorption coefficient evaluated by the same method as for the allowed transition leads to Equation 27.54.

(27.9) **Allowed indirect transitions**

Equation 27.56 gives the absorption coefficients for allowed indirect optical transitions. These transitions are of high importance for silicon and germanium based solar cells and photodetectors. The derivation is somewhat involved and may be found in *Wave Mechanics of Crystalline Solids* by R. A. Smith, Chapter 13. Locate and read this section in the textbook. Note specially the use of second order perturbation theory in the derivation, and appreciate the similarity of how electron-phonon and electron-photon interactions are treated on an equal footing. Discuss why photon-phonon interactions are left out and what sort of impact should such interactions have on the observed absorption spectra.

(27.10) **Exciton matrix elements**

(a) Verify the results of Table 27.2 for excitons by using the results of the wavefunction and eigenvalues of hydrogen atoms suitably modified for the exciton.

(b) One of the steps in the exciton model is to relate $\langle f|\hat{\mathbf{p}}|i\rangle$ to the value of the hydrogen atom wavefunction $\psi(0)$ at the origin. The following is a brief outline – for a detailed derivation, please refer to Chapter 14 of *Physics of Photonic Devices* by S. L. Chuang. The exciton wavefunction is written as a linear combination $\psi(\mathbf{r}_c, \mathbf{r}_v) = \sum A_{\mathbf{k}_c,\mathbf{k}_v} \psi_{\mathbf{k}_v}(\mathbf{r}_v) \psi_{\mathbf{k}_c}(\mathbf{r}_c)$ in Equation 27.57 in this chapter, which in the electron-hole center of mass frame yields the exciton eigenfunctions and eigenvalues in Equation 27.58. Show by following these steps that $\langle f|\hat{\mathbf{p}}|i\rangle$ is proportional to the sum over $\mathbf{k}$ of the coefficients $A_{\mathbf{k}_c=+\mathbf{k},\mathbf{k}_v=-\mathbf{k}}$. Then, use the standard Fourier result analogous to $g(x) = \sum C_k e^{ikx} \implies \sum C_k = g(0)$, which means the sum of the Fourier coefficients is the value of the function at the origin.

(27.11) **Electroreflectance oscillations**

A formula for the Franz–Keldysh oscillations in the **reflectivity** spectra from a semiconductor is obtained from the asymptotic form of the Airy function, given by $\Delta R(\hbar\omega)/R \propto e^{-\frac{2\Gamma\sqrt{\hbar\omega - E_g}}{(\hbar\omega_F)^{\frac{3}{2}}}} \cos[\frac{4}{3}(\frac{\hbar\omega - E_g}{\hbar\omega_F})^{\frac{3}{2}} + \phi] \frac{1}{(\hbar\omega)^2(\hbar\omega - E_g)}$, where $\hbar\omega_F$ is described in Equation 27.63, Γ is a linewdith of the transition, and ϕ is a phase angle.

(a) Justify how this may result from the absorption coefficient due to Franz–Keldysh oscillations described in Equation 27.63.

(b) Use this equation for Franz–Keldysh oscillations in the reflectivity spectra to fit and extract the electric field from Fig. 27.16. Note that these oscillations are for photon energies larger than the bulk bandgap energy; compare with the absorption curves in Fig. 27.15.

(c) Check your field with those extracted in the reference for Fig. 27.16.

Heavenly Light: Solar Cells and Photodetectors

28

28.1 Solar sells and photodetectors: a résumé 753
28.2 Solar cells 754
28.3 Shockley–Ramo theorem 763
28.4 Photodetectors 763
28.5 Avalanche photodiodes 766
28.6 Quantum well infrared photodetectors 766
28.7 Electro-absorption modulators 767
28.8 Solar blind photodetectors 767
28.9 Chapter summary section 768
Further reading 768
Exercises 768

In the last chapter, we explored a range of processes by which photons are absorbed by semiconductors. This process of absorption is the basis for a range of photonic devices such as solar cells, photodetectors, and modulators. In this chapter, we investigate:

- How is light energy converted to electrical energy by semiconductor solar cells, and what are the fundamental limits of their efficiencies?
- How are optical absorption processes in semiconductors used to make efficient photodetectors, and over what ranges of wavelengths, timescales, and efficiencies can they operate?
- What is the operating principle of semiconductor based optical modulators, and what are their limitations?

28.1 Solar sells and photodetectors: a résumé

Right after the realization of the first silicon pn junctions, Pearson et al. at Bell Labs observed that shining light on the junction region produced a large voltage across electrodes connected to the p- and n-regions of the junction. This was the first silicon solar cell. The operation mechanism was tracked to the absorption of photons by electrons in the valence band that were subsequently excited to the conduction band. The built-in electric field in the junction swept out carriers of opposite type to the two electrodes, producing the photovoltage.

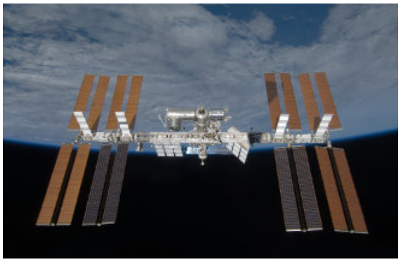


Fig. 28.1 Arrays of solar cell panels on the International Space Station.

The initial findings led to activity in determining how efficient solar cells could potentially be (as **solid-state engines** with no mechanical moving parts) in converting solar radiation energy into useful electrical energy. As semiconductors other than silicon (such as CdS and others) were explored for solar cells, the efficiency was found to be dependent on the energy bandgap. Several works identified the optimal bandgap for the highest efficiency, accounting for losses at contacts, series resistances, reflection, etc. In 1961, the **Shockley–Queisser limit** of solar cells was introduced based purely on thermodynamic and quantum mechanical considerations by the two scientists. The efficiencies of today's solar cells installed in rooftops, terrestrial solar farms, and in space vehicles (Fig. 28.1) get close to, but fall short of this limit.

The efficiency of solar cells have increased by increasing perfection of the choices of semiconductors and their junction design, methods to ensure the maximal absorption of sunlight, by increasing the concentration of sunlight using concentrator cells, and by the use of cascaded junctions of several bandgaps. Several semiconductor materials have been introduced, such as crystalline silicon (c-Si) for high efficiency, and amorphous silicon (a-Si) for large-area and low cost cells, III-V semiconductors such as GaAs and InGaP/GaAs/Ge multijunction cells, II-VI semiconductors such as CdTe and CdS, halide perovskite semiconductors, organic semiconductors, and even dye-sensitized solution-based solar cells. The maximum measured efficiencies for single-junction solar cells are $\lesssim 30\%$, and for multijunction cells are $\lesssim 50\%$. They are discussed later in this chapter.

Photon absorption in semiconductors and conversion into either a photocurrent or photovoltage is also the basis for the operation of **photodetectors** for information processing. Semiconductor photodetectors typically use interband transitions and can operate in the photoconductive mode, in the photovoltaic mode, or even in the avalanche mode with varying levels of sensitivity and speed. Semiconductor photodetector arrays are used today for high-speed and high-resolution imaging in myriad applications. The ability to allow or forbid the absorption of photons in semiconductor structures with external electric fields (or by current injection) forms the basis of optical modulators, which are central in high-speed communication systems. The principles of operation of semiconductor photodetectors and modulators are described in this chapter, following the discussion of solar cells.

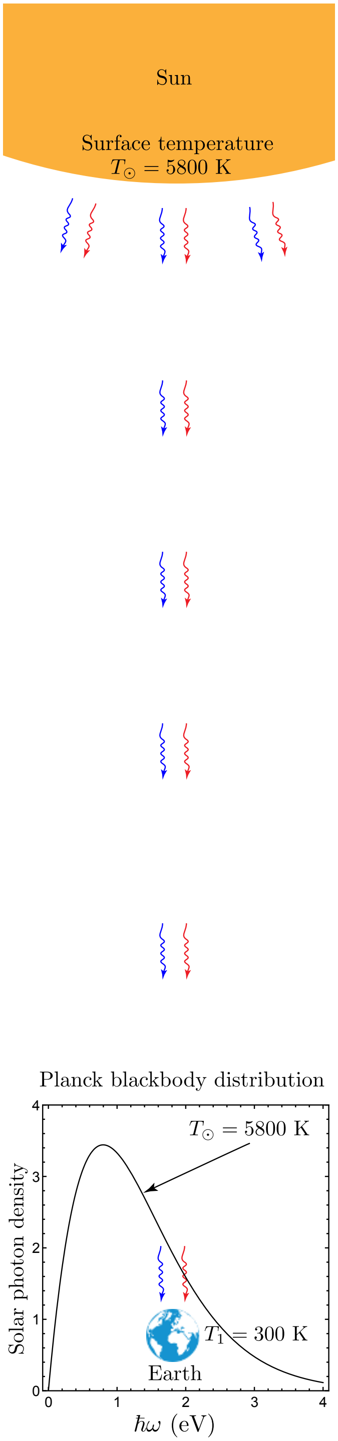


Fig. 28.2 Light from the Sun.

28.2 Solar cells

Every second, the Sun loses $\Delta m \approx 4 \times 10^3$ kg of its total mass $M_\odot = 2 \times 10^{30}$ kg due to the fusion reaction of 4×10^{11} kg of hydrogen atoms to produce helium (H + H → He + energy). The mass lost is radiated as electromagnetic waves following Einstein's relation $E = (\Delta m)c^2 \approx 4 \times 10^{20}$ J, where c is the speed of light. The projected life of the Sun exceeds 10^{10} years. The photons are radiated in all directions. The temperature of the Sun surface is $T_\odot = 5800$ K, and the temperature on the surface of the Earth is $T_1 = 300$ K, as shown in Fig. 28.2. Thus, the spectrum of photon energies that reach the outside of Earth's atmosphere closely resembles Planck's blackbody radiation pattern of characteristic temperature $T_\odot$ at the surface of the Sun, as shown in Fig. 28.2. The photons from the Sun reach outside the Earth's atmosphere roughly 8 minutes after being emitted from the Sun's surface, by traveling ballistically through space. They are thus in thermodynamic equilibrium with the surface of the Sun, like electrons injected from the FET source are in equilibrium with the source (Chapter 19).

Fig. 28.3 shows how photons from the Sun are incident on a semiconductor pn junction solar cell, with the resulting energy band di-

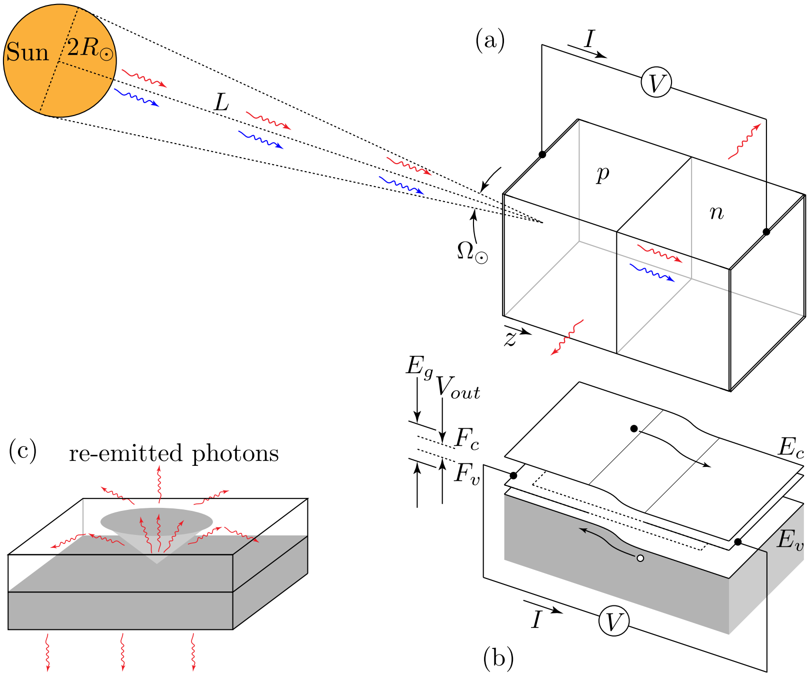


Fig. 28.3 (a) A semiconductor p-n junction solar cell receiving radiation from the Sun, and generating a photovoltage and photocurrent. Solar radiation within the solid angle $\Omega_\odot = 6.8 \times 10^{-5}$ steradian reaches the cell, making it a fraction $f_\odot = \Omega_\odot / 4\pi$. (b) The energy band diagram of the solar cell in operation. (c) The cell also re-emits photons. The maximum efficiency depends on maximizing the absorption and conversion, and minimizing the re-radiation.

agram and the fact that the semiconductor can also re-emit photons after absorbing them. We will use this figure to develop a model for the output characteristics and efficiency of solar cells. Before getting into the details, it is important to note that the accurate spectrum of photons incident on solar cells is shown in Fig. 28.4. We assume the blackbody radiation spectrum at $T_\odot$ in the following discussion for simplicity. The number of photons reaching the solar cell per unit energy per unit area per unit time is obtained from Planck's formula $N_{ph} = \frac{1}{\pi^2 \hbar^3 c^3} \cdot \frac{(\hbar\omega)^3}{e^{\frac{\hbar\omega}{k_b T_\odot}} - 1}$ using $f_\odot$ defined in Fig. 28.3:

$$\frac{dn_{ph}}{d(\hbar\omega)} = \frac{f_\odot N_{ph} c}{\hbar\omega} = \frac{f_\odot}{\pi^2 \hbar^3 c^2} \cdot \frac{(\hbar\omega)^2}{e^{\frac{\hbar\omega}{k_b T_\odot}} - 1}, \tag{28.1}$$

where $\hbar\omega$ is the photon energy. For a pn junction made of a semiconductor of bandgap E_g, we assume that photons of energy $\hbar\omega < E_g$ are not absorbed, and all photons of energy $\hbar\omega \geq E_g$ are absorbed. The photon density n_{ph} absorbed by the solar cell per unit area per unit time is then obtained by integrating Equation 28.1 for $\hbar\omega \geq E_g$:

$$n_{ph}(E_g) = \int_{E_g}^{\infty} \frac{dn_{ph}}{d(\hbar\omega)} \cdot d(\hbar\omega) = \frac{f_\odot (k_b T_\odot)^3}{\pi^2 \hbar^3 c^2} \int_{u_g}^{\infty} \frac{u^2 du}{e^u - 1}, \tag{28.2}$$

where $u = \hbar\omega / k_b T_\odot$ and $u_g = E_g / k_b T_\odot$. This density is of the order 10^{17} cm^{-2}s^{-1}, and is a measure of the net sheet density of electron-hole pairs produced by the absorbed photons in the semiconductor

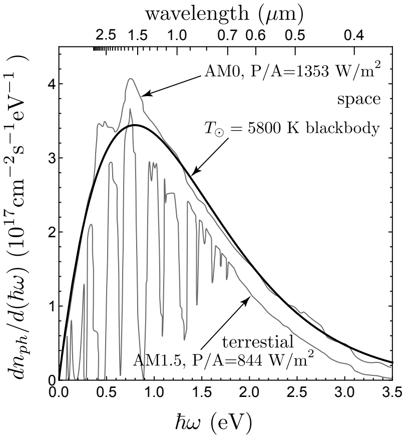


Fig. 28.4 Solar spectra at AM0 and AM1.5 compared to Planck's blackbody radiation formula at $T_\odot = 5800$ K. AM0 stands for atmospheric mass 0. It is the spectrum just outside the Earth's atmosphere, and is nearly identical to Planck blackbody radiation formula. AM1.5 is the spectrum that makes it to the surface of the Earth after several sharp absorptions and uniform attenuation in the atmosphere. AM0 is the spectrum incident on solar cells on satellites, and AM1.5 on solar cells on the surface of the Earth.

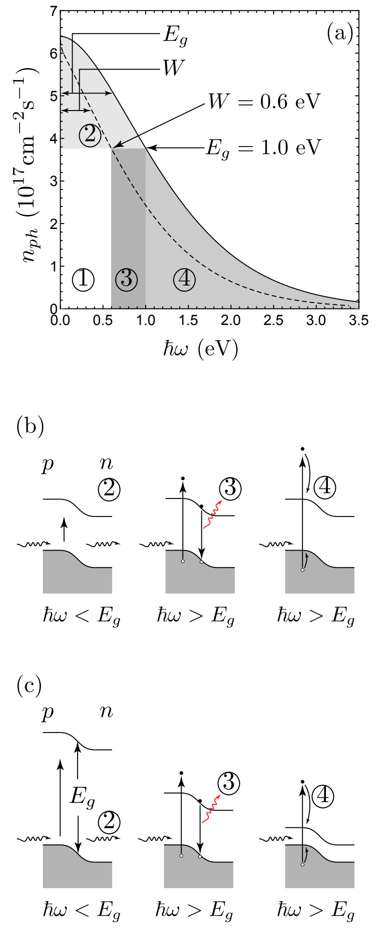


Fig. 28.5 (a) Photon density vs. energy, and (b, c) energy band diagrams indicating the three sources of loss of the solar cell. The losses for the same bandgap semiconductor pn junction are shown in (b) for increasing photon energy. The losses for the same photon energy for decreasing semiconductor bandgap are shown in (c). This method of pictorial analysis of solar cells in (a) was introduced by C. H. Henry in J. Appl. Phys., **51**, 4495 (1980). Our treatment closely follows this paper.

per unit time. It is shown as a function of energy in Fig. 28.5 as the solid line. Therefore, the net power of solar radiation incident on the semiconductor per unit area is given by $P_{in} = \int_0^\infty d(\hbar\omega) \cdot \hbar\omega \cdot \frac{dn_{ph}}{d(\hbar\omega)}$, or alternatively the area under the solid curve in Fig. 28.5. For the blackbody spectrum approximation, the value of this integral is ~ 1325 W/m^2. This is close to the true AM0 value of 1353 W/m^2 as indicated in Fig. 28.4, but higher than the AM1.5 direct spectrum terrestrial value of 844 W/m^2. The order of magnitude of incident solar power is thus of the order of 1 kW/m^2.

Choice of the semiconductor for a single pn junction solar cell fixes the energy bandgap E_g, which is a point on the solid line in Fig. 28.5. For example, consider a bandgap of $E_g = 1.0$ eV as shown. Of the total incident solar power, corresponding to this choice of bandgap, the portions shown as the shaded areas ②, ③ and ④ are lost, and only the **Area** ① is successfully converted into electrical power P_{out} that is delivered to the external circuit by the solar cell.

Area ② is lost because photons of energy smaller than the bandgap are not absorbed by the solar cell as indicated in the leftmost energy band diagram in Fig. 28.5.

Area ③ is lost because the semiconductor can re-emit photons by radiative recombination of electron-hole pairs (middle energy band diagram), or lose them to non-radiative recombination. We assume radiative recombination in our analysis.

Area ④ is lost because when the energy of the incident photon is larger than the bandgap of the semiconductor, the excited electron relaxes to the bottom of the conduction band, and the hole relaxes to the top of the valence band, both by the emission of phonons (right energy band diagram). Thus $\hbar\omega - E_g$ portion of energy is lost as heat.

Fig. 28.5 also shows a dashed line whose length from the y-axis is the useful work W delivered by an optimally biased pn junction, as will be discussed shortly. The work W connects to thermodynamic arguments, which we now invoke to explain the efficiency of the solar cell. The area of region ① is $P_{out} = W \cdot n_{ph}(E_g)$, and the efficiency of the solar cell is given by $\eta = P_{out}/P_{in}$.

Before delving into the details, the design criteria for efficient solar cells can be already seen in Fig. 28.5. The goal is to maximize area ①, the rectangle whose diagonal starts at the origin and terminates on the dashed line. For a single-junction solar cell of bandgap E_g, a horizontal line is drawn from the solid line to the y-axis. The intercept of this line with the dashed "work" line is the second vertex of rectangle ①. For example, for a semiconductor of bandgap $E_g = 1.0$ eV, $n_{ph} \sim 3.75 \times 10^{17}$ cm^{-2}s^{-1}, and we will show shortly that the maximum work is $W = 0.6$ eV, defining the area ①. If the bandgap of the semiconductor is chosen to be too small, the rectangle ① is tall and very thin: most photons are absorbed, but most of the energy is lost to heat. If the bandgap is too large, rectangle ① is short and fat: most photons pass through unabsorbed. The goal then is to find the optimal semiconductors with bandgaps that maximize area ①.

Consider Fig. 28.3 (a). The net current density $J = I/A$ that flows in the external circuit is due to the electron-hole pairs created by the solar light $J_{ph} = qn_{ph}$, minus J_{rad}, the portion due to electron-hole pairs that are lost in recombination that produces spontaneous emission. Equation 28.2 gives n_{ph}, from which $J_{ph} = qn_{ph}$ is evaluated. Therefore finding J_{rad} at an applied voltage V will allow us to find the characteristics of the solar cell. These calculated values are shown in Fig. 28.6. We first quantify the physics that lead to these plots.

J_{rad} is the current density of the solar cell behaving as a light-emitting diode (LED). Note that for an efficient solar cell J_{rad} must be minimized, whereas for a LED it must be maximized. The LED will be discussed in Chapter 29 from the point of view of Einstein's model of spontaneous and stimulated emission. Here we use a thermodynamic argument to evaluate, as a function of the applied voltage V, the spontaneously emitted steady state photon occupation function $\bar{n}_\omega$, the resulting steady state emitted photon spectra[1] $\bar{\rho}$, and J_{rad}.

When a forward–bias voltage V is applied across the solar cell, the quasi-Fermi level F_c of electrons in the CB on the n-side and F_v of electrons in the VB on the p-side split according to the relation $F_c - F_v = qV$, as discussed in Chapter 18. F_c and F_v represent the free energy needed to add an electron to the conduction band or the valence band respectively. The work done by carriers in the solar cell when an electron drops from F_c to F_v to spontaneously emit a photon is[2] $\Delta W_s = F_c - F_v = qV$. From the first law of thermodynamics (Chapter 4), energy conservation requires this to satisfy $\Delta U_s = \Delta W_s + \Delta Q_s$, where ΔU_s is the energy, and $\Delta Q_s = T_1 \Delta S_s$ is the entropy, both of which are lost by the semiconductor. By energy conservation, $\Delta U_s + \Delta U_e = 0$, where ΔU_e is the energy gained by the electromagnetic modes denoted by the subscript e. Since the semiconductor is in thermal equilibrium with the electromagnetic modes it is radiating into, the total entropy of the semiconductor+electromagnetic system is already maximized (Chapter 21) and cannot change. This implies $T_1 \Delta S_s = T_1 \Delta S_e$.

The total entropy S_e in the photon modes and their total energy U_e is (Chapter 4, Equation 4.4 for bosons):

$$S_e = -k_b \sum_{\omega'} [\bar{n}_{\omega'} \ln \bar{n}_{\omega'} - (\bar{n}_{\omega'} + 1) \ln(\bar{n}_{\omega'} + 1)] \text{ and } U_e = \sum_{\omega'} \hbar\omega' \cdot \bar{n}_{\omega'} \tag{28.3}$$

where $\bar{n}_{\omega'}$ is the steady-state occupation number of photons in the mode of energy $\hbar\omega'$. Spontaneous emission of a *single photon* from the solar cell leads to $\sum \Delta n_{\omega'} = +1$. Differentiating Equation 28.3 and using $\sum \Delta n_{\omega'} = +1$ gives the energy and entropy increase in the photon modes as

$$\Delta S_e = k_b \ln \frac{\bar{n}_\omega + 1}{\bar{n}_\omega} \text{ and } \Delta U_e = \hbar\omega. \tag{28.4}$$

Since these are equal to the energy and entropy lost by the semiconductor, we equate $\Delta U_e - T_1 \Delta S_e = \Delta U_s - T_1 \Delta S_s = qV$. This immediately yields the steady-state photon occupation number $\bar{n}_\omega$ as a function of the applied voltage:

---

[1] We expect $\bar{n}_\omega$ and $\bar{\rho}$ to be different from the blackbody radiation spectrum only for $V \neq 0$, but merge with the blackbody spectrum at $V = 0$. The $V = 0$ radiation pattern should be identical to the blackbody spectrum for $T = T_1$, the temperature of the *semiconductor*, and not the temperature $T_\odot$ of the Sun. This is because this part of the radiation is in thermal equilibrium with the semiconductor.

[2] This is explicitly written as $\Delta W_s = [F_c \cdot (N_n + 1) + F_v \cdot (N_p)] - [F_c \cdot (N_n) + F_v \cdot (N_p + 1)] = F_c - F_v = qV$, where N_n and N_p are the particle numbers at the two quasi-Fermi levels.

$$qV = \Delta U_e - T_1 \Delta S_e = \hbar\omega - k_b T_1 \ln\left(\frac{\bar{n}_\omega + 1}{\bar{n}_\omega}\right) \implies \boxed{\bar{n}_\omega = \frac{1}{e^{\frac{\hbar\omega - qV}{k_b T_1}} - 1}}. \tag{28.5}$$

Note that the occupation function of photons $\bar{n}_\omega$ radiated by the solar cell looks deceptively like the blackbody spectrum at temperature T_1, except for the voltage V appearing in the exponential. Writing $v_g = \frac{d\omega}{dk} = \frac{c}{n_r}$ as the group velocity of light, $k = \frac{n_r \omega}{c}$ where n_r is the refractive index of the semiconductor and c is the free-space light speed, the net radiative *photonic* current density J^{ph}_{rad} is obtained by summing over all k-modes the semiconductor can emit into:

$$J^{ph}_{rad} = \frac{2}{L^3} \sum_k v_g(k) \cos\theta \bar{n}_\omega = \frac{2}{L^3} \int \frac{k^2 \sin\theta dk d\theta d\phi}{(\frac{2\pi}{L})^3} \cos\theta v_g(k) \bar{n}_\omega, \tag{28.6}$$

where Fig. 28.3 (c) indicates the angles into which the re-emitted photons go. In the upper half light emitted into a cone defined by $\sin\theta_{cr} \leq \frac{1}{n_r}$ escape the semiconductor pn junction, whereas those at higher angles undergo total internal reflection. All photons emitted into the lower half $\frac{\pi}{2} \leq \theta \leq \pi$ are assumed to be absorbed by the substrate on which the solar cell is deposited. The semiconductor is assumed to emit photons of energy $\hbar\omega \geq E_g = \hbar\omega_g = \hbar(\frac{c}{n_r} k_g)$. Then,

$$J^{ph}_{rad} = \frac{c}{2\pi^2 n_r} \left(\int_0^{\theta_{cr}} d\theta \cdot \sin\theta \cdot \cos\theta + \int_{\frac{\pi}{2}}^{\pi} d\theta \cdot \sin\theta \cdot \cos\theta \right) \cdot \int_{k_g}^{\infty} k^2 dk \cdot \bar{n}_\omega$$

$$= \frac{(1 + n_r^2)}{4\pi^2 c^2} \int_{\omega_g}^{\infty} \frac{\omega^2 d\omega}{e^{\frac{\hbar\omega - qV}{k_b T_1}} - 1} \implies \boxed{J_{rad} = \frac{q(1 + n_r^2)(k_b T_1)^3}{4\pi^2 c^2 \hbar^3} \int_{u_g}^{\infty} \frac{u^2 du}{e^{u-v} - 1}}, \tag{28.7}$$

where $u_g = \frac{E_g}{k_b T_1}$ and $v = \frac{qV}{k_b T_1}$ are used, along with $J_{rad} = qJ^{ph}_{rad}$, since for each photon emitted, an electron-hole pair is lost. Note that J_{rad} depends explicitly on the applied voltage V. The net charge current flowing in the diode must be

$$J = [\underbrace{J_{rad}(V) - J_{rad}(V=0)}_{=\text{diode dark current } "J_{dark} = J_0(e^{qV/k_b T_1} - 1)"}] - \underbrace{qn_{ph}}_{\text{solar}}, \tag{28.8}$$

where the first bracketed term on the right is the ideal pn diode current with 100% radiative transitions that is characterized by the temperature of the semiconductor T_1. The term qn_{ph} (from Equation 28.2) is the photocurrent density due to the Sun characterized by temperature $T_\odot$. Figures 28.6 and 28.7, that characterize the operation and efficiency of the solar cell are evaluated using Equation 28.8. Before describing approximations for analytical evaluation, let us discuss these exact results.

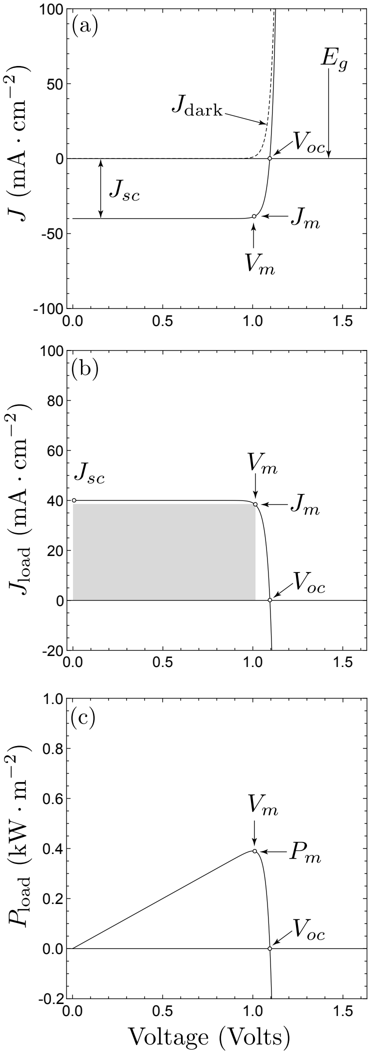


Fig. 28.6 Solar cell J-V curves and output power vs. applied voltage for a single-pn junction of energy bandgap E_g = 1.42 eV. (a) The dashed line shows the dark current density and the solid line the current density with solar radiation AM0. The open-circuit voltage V_{oc} and the short-circuit current J_{sc} are labeled. (J_m, V_m) are the coordinates of the bias point where the area $J \cdot V$ is maximized. (b) The current delivered to the load, same as (a) but inverted. (c) The power delivered to the load. Note that for AM0 solar radiation, P_{in} = 1.353 kW·m^{-2}, and the maximum power delivered for this cell is ∼ 0.4 kW·m^{-2}, which is a $\eta = \frac{P_{load}}{P_{in}} \sim 30$ %.

Fig. 28.6 (a) shows the dark current and the current at AM0 solar illumination for a semiconductor such as GaAs of bandgap 1.42 eV. The net photocurrent density is ∼ 40 mA·cm^{-2}, a typical value for this radiation intensity. Since this current is negative while the applied voltage is positive, $P = J \cdot V$ is negative, which means the cell is *producing* power, not dissipating it.

Because this current flows even when the outside terminals (Fig. 28.3) are shorted to make $V = 0$, the y-axis intercept is called the **short-circuit current** J_{sc}. If the terminals are left open, no current can flow, and $J = 0$ then defines the **open-circuit voltage** V_{oc}. Note that qV_{oc} is a few 100 meV below the semiconductor bandgap E_g.

Fig. 28.6 (b) is identical to (a), but is it's negative – this is what is delivered to the load: $J_{load} = -J$. The solar cell delivers power $P_{load} = J \cdot V$ to the load as shown in Fig. 28.6 (c) for various voltage biases. It is clear that there is an optimal bias voltage at which the power delivered is maximized: this voltage is denoted V_m and the corresponding current as J_m, and the maximum power as $P_m = J_m \cdot V_m$. This is clearly the area of the shaded rectangle in Fig. 28.6 (b), which is a fraction of the slightly bigger rectangle of area $J_{sc} \cdot V_{oc}$. The ratio of the areas is called the **fill factor** $= \frac{J_m \cdot V_m}{J_{sc} \cdot V_{oc}}$, which is often expressed as a percentage. Note from Fig. 28.6 (c) that the optimal bias point is just below V_{oc}, and the maximum power delivered for a bandgap of $E_g = 1.42$ eV is ∼ 0.4 kW·m^{-2}, about 30% of AM0 incident power.

Fig. 28.7 (a) shows how this maximum delivered power changes as a function of the bandgap of the semiconductor used for the solar cell. Unfortunately one cannot extract more than about 0.4 kW·cm^{-2}, and the maximum is around $E_g = 1.4$ eV. The corresponding V_{oc} and the work done per absorbed photon $W_m = J_m V_m / e n_{ph}$ are shown in Fig. 28.7 (b) as a function of the semiconductor bandgap. Since for AM0 $P_{in} = \int_0^\infty d(\hbar\omega) \cdot \hbar\omega \cdot \frac{dn_{ph}}{d(\hbar\omega)}$, the maximum efficiency of the single-junction solar cell $\eta = \frac{J_m V_m}{P_{in}}$ is plotted as a function of the bandgap in Fig. 28.7 (c). This is called the **Shockley–Queisser** limit after the researchers who first derived it in 1961 based on detailed balance in thermodynamics (see Figures 18.16 and 28.8).

We now discuss a few approximations that can be made to simplify the above calculations. Since $E_g >> k_b T_1$ and $qV >> k_b T_1$ are satisfied near the optimal bias condition, the $J_{rad}(V = 0)$ component of the current of Equation 28.8 is negligible, and

$$J_{load} \approx qn_{ph} - \frac{q(1+n_r^2)(k_bT_1)^3}{4\pi^2c^2\hbar^3} e^v \underbrace{\int_{u_g}^{\infty} e^{-u}u^2du}_{\approx u_g^2 e^{-u_g}} \approx qn_{ph} - J_0 e^{\frac{qV-E_g}{k_bT_1}}, \tag{28.9}$$

where $J_0 = \frac{q(1+n_r^2)E_g^2(k_bT_1)}{4\pi^2c^2\hbar^3}$ is a characteristic current density dependent on the semiconductor bandgap, refractive index, and its temperature. The analytical approximation of the open circuit voltage V_{oc} is then

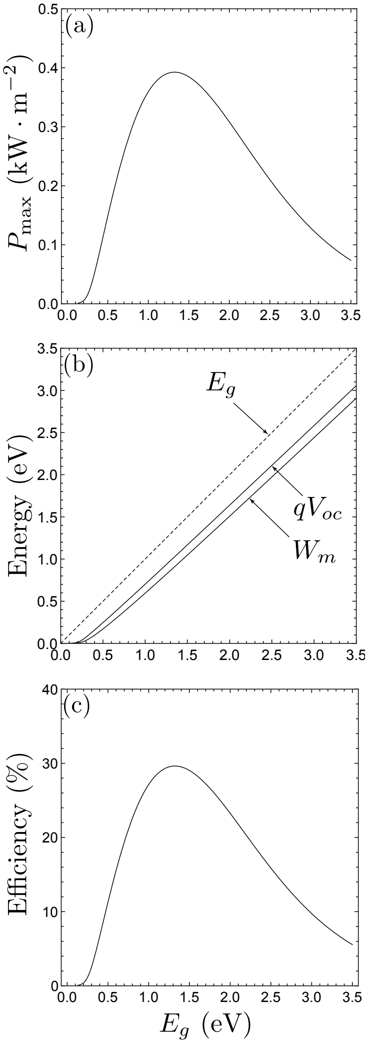


Fig. 28.7 (a) Maximum output power of a single-bandgap pn junction solar cell at AM0 solar flux as a function of the semiconductor bandgap E_g. (b) The open circuit voltage and the work done per absorbed photon $W_m = J_m V_m / e n_{ph}$ are typically ∼ 0.5 eV less than the bandgap, and (c) the maximum efficiency as a function of E_g.

obtained when $J_{load} = 0$:

$$qV_{oc} \approx E_g - k_b T_1 \ln(\frac{J_0}{qn_{ph}}), \tag{28.10}$$

and the maximum output power is obtained from $d(J_{load} \cdot V)/dV = 0$, which gives V_m, J_m, and the work done per absorbed photon W_m as

$$qV_m \approx qV_{oc} - k_b T_1 \ln(1 + \frac{qV_m}{k_b T_1}), \qquad J_m = \frac{qn_{ph}}{1 + \frac{k_b T_1}{qV_m}}, \text{ and}$$

$$W_m \approx E_g - k_b T_1 [\ln(\frac{J_0}{qn_{ph}}) + \ln(1 + \frac{qV_m}{k_b T_1}) + 1], \tag{28.11}$$

where the first equation must be solved numerically. Using the parameters used for the exact evaluation of Figures 28.6 and 28.7, it can now be verified that as long as $E_g >> k_b T_1$, these approximations are actually excellent (see Exercise 28.3).

Fig. 28.8 Hans Joachim Queisser, with Shockley, established the efficiency limits on solar cells and made early contributions to the development of semiconductor LEDs.

Fig. 28.9 Charles Henry, a pioneer who realized the importance of quantum well LEDs and lasers and the analogy of electrons and holes in quantum wells to light propagation in waveguides. He wrote the influential paper on the design of multijunction solar cells to beat the single-junction Shockley–Queisser limits, which we discuss here.

We now investigate how higher efficiencies than the single-junction Shockley–Queisser limit can be obtained by the clever use of **multjunction solar cells**. The strategy of such cells is straightforward: recover as much of lost areas ②, ③, and ④ in Fig. 28.5 (a) by the use of stacked pn junctions of different bandgaps and maximize area ①, the power delivered to the load. Fig. 28.10 illustrates how this is achieved, and Fig. 28.11 shows an example semiconductor material stack for a triple-junction cell.

Fig. 28.10 (a) for the single junction schematically shows the single junction output with the energy band diagram in which the optimal voltage V_m is indicated, and the shaded rectangle is the power delivered to the load. Fig. 28.10 (b) shows the case when two pn junctions are stacked: light is incident on the surface side of the wider bandgap E_{g1}. Photons of energy $\hbar\omega > E_{g1}$ above the bandgap generate a photovoltage and photocurrent in junction 1. Photons of energy $\hbar\omega < E_{g1}$ pass through to junction 2 of bandgap E_{g2} and generate a photovoltage and photocurrent in junction 2. The junctions are connected in series using interband tunnel junctions (Chapter 24) between them such that the valence band of the wider bandgap p-side is aligned with the conduction band edge of the narrower bandgap n-side. This ensures that the holes in the wider-bandgap p-side come in equilibrium with the electrons in the narrower bandgap n-side and share the same Fermi level. Because of these conditions, the current through the two cells is identical and equal to $J_m \approx qn_{ph}$ corresponding to the chosen E_{g1}, E_{g2}.

Thus it is necessary to design the cells such that they are **current-matched**, i. e., their $J_m = qn_{ph}$ are equal. Pictorially, the step heights of the J_m steps in Fig. 28.10 (b) then need to be identical. Furthermore, the voltages generated in the two cells *add*, leading to a higher open circuit voltage and thus a higher V_m than a single-junction cell when designed and biased properly[3]. Therefore, a multijunction cell will have a smaller J_{sc} and a larger V_m than a single-junction cell, such that the product $J_m V_m$ is larger than the single-junction cell. This is

[3] The multijunction design discussed here is the most commonly used type. Other possibilities (such as voltage matched junctions) or different ways of combining various junctions are also in use depending on the semiconductor materials and feasibility.

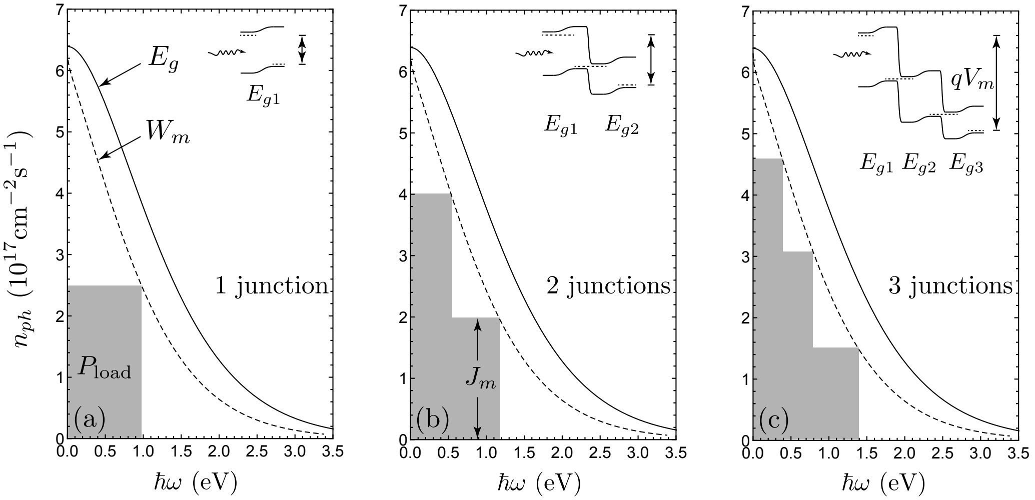


Fig. 28.10 Graphical design of multijunction solar cells for maximum efficiencies shown for AM0 incident power (not exact, meant for illustration). The maximum power delivered to the load indicated by the shaded area can be increased by the use of multijunction (or tandem) cells. The cells, or pn junctions are connected in series such that the net photocurrent J_m, proportional to the height of the rectangles are the same for each junction of the multijunction cells, whereas the quasi-Fermi level differences add, leading to a larger voltage delivered to the load. (a) Single junction, (b) 2 junctions, and (c) 3 junction solar cells, with schematic energy band diagrams that show each stacked pn diode connected by tunnel junctions.

pictorially represented by the fact that the total area of the shaded rectangles in Fig. 28.10 (b) is larger than that in Fig. 28.10 (a).

The bandgaps and the currents needed are identified by finding the maximum area possible while satisfying the above current matching requirements. This typically requires numerical optimization. The same principle can be carried over to more junctions in tandem cells, as indicated in Fig. 28.10 (c) for three junctions with three bandgaps, that in turn covers a larger area than a two junction cell.

Since the area under the dashed line is the maximum power that can be delivered to the load, it may be approached by using a large number of junctions of different bandgaps. For multijunction cells, it is more cost-effective if the solar radiation is **concentrated** to higher than the standard radiation to make use of smaller areas of the cells. This is typically done by arrangements of lenses that focus the light on the solar cell. For AM1.5 solar intensity, it is found that increasing the concentration of photons by a factor of 1000 leads to theoretical efficiencies of $\eta_1 \approx 37\%, \eta_2 \approx 50\%, \eta_3 \approx 56\%$, and $\eta_{36} \approx 72\%$ where the subscripts indicate the number of bandgaps used in the multijunction cell. How do these predictions compare with experiment?

Fig. 28.11 shows an example of the bandgaps used in triple junction cells. The base, narrowest bandgap layer is Ge, and the larger bandgap layers chosen are InGaAs (or GaAs) and InGaP because they are almost lattice-matched to Ge and can be grown epitaxially without too many non-radiative defects. Furthermore, it is critical to be able to connect each bandgap junction to the other by very low resistance

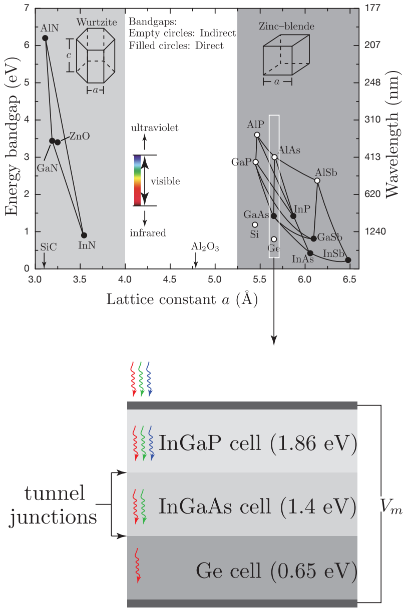

AlN
Wurtzite
Bandgaps:
Empty circles: Indirect
Filled circles: Direct
Zinc-blende
ZnO
GaN
InN
SiC
ultraviolet
visible
infrared
Al_2O_3
AlP
AlAs
GaP
AlSb
InP
GaAs
Si
Ge
GaSb
InAs
InSb
Energy bandgap (eV)
Lattice constant a (Å)
Wavelength (nm)
177
207
248
310
413
620
1240
InGaP cell (1.86 eV)
InGaAs cell (1.4 eV)
Ge cell (0.65 eV)
tunnel junctions
V_m

Fig. 28.11 Design of a triple junction solar cell: choice of semiconductor materials. Typically lattice-matched structures prevent formation of defects.

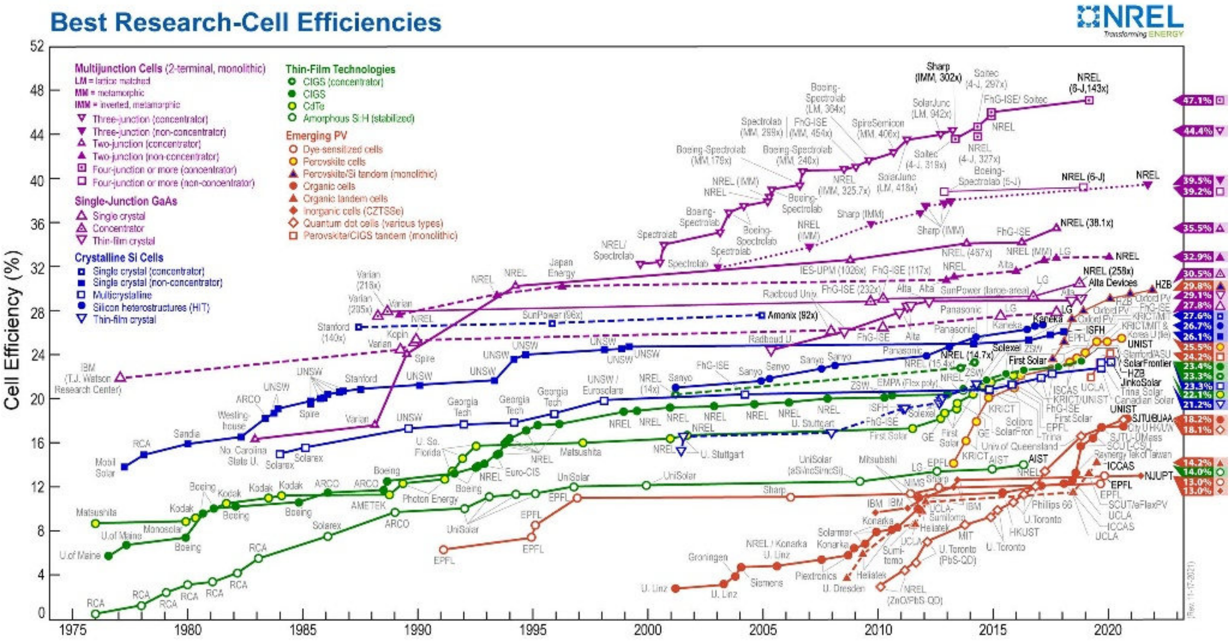


Fig. 28.12 Evolution of solar cell efficiencies over the years for various semiconductor material platforms as of January 2022. This plot is courtesy of the National Renewable Energy Laboratory, Golden, CO.

tunnel-junctions. These have been achieved successfully for the layer structure shown. Fig. 28.12 shows the highest efficiencies of solar cells achieved to date. Three-junction (concentrator) cells have reached ~ 44 %, and six-junction cells have reached ~ 47 %. These are lower than the calculated maximum values. But note that the calculations above have neglected several sources of loss.

Let us discuss the sources of loss. There are a host of extrinsic losses such as reflection of sunlight (which is reduced by anti-reflection coatings), the ohmic contacts shadow part of the sunlight, the series resistances in the conductive regions and contacts, loss in tunnel junction resistances, and in general absorption and recombination processes that are undesirable for the performance. Non-radiative recombination is a source of losses. Typically lattice-matched structures (such as Ge/InGaAs/InGaP) prevent formation of defects such as dislocations that cause non-radiative recombination and current leakage. If defects cannot be prevented, it must be ensured that they are carefully placed so as to not compromise the efficiency. Sometimes concentrating to too high values degrades the crystalline quality by heating up the cell. These losses account for the difference between the calculated efficiency and what is obtained ($\eta_3 \sim 56$ % vs. η_3(expt.) ~ 44 %).

It is interesting to consider the efficiency of a **Carnot heat engine** tasked to extract work out of a reservoir at temperature $T_\odot$ and dumping heat into a reservoir at temperature T_1. In that case, the maximum efficiency attainable is $\eta_{\text{Carnot}} = 1 - \frac{T_1}{T_\odot} \sim 95\%$. The maximum efficiency of *solar* energy conversion however is $\eta_{\max} = 1 - \frac{4}{3}\frac{T_1}{T_\odot} \sim 93\%$. Both these values are much higher than the ~ 72 % achieved with multijunction solar cells (in theory!). Exercise 28.4 discusses these limits and provides references to find where the remaining ~ 20 % is lost.

28.3 Shockley–Ramo theorem

Before discussing the properties of semiconductor photodetectors, it is useful to preemptively dispel a source of confusion that often arises in understanding the time-dependent response of such devices. Consider Fig. 28.13 (a), where an electron is traveling in vacuum from the left to the right electrode at a velocity $\mathbf{v}(t)$ at time t. What is the current $I(t)$ flowing in the external circuit *at that time instant*? The **Shockley–Ramo theorem** asserts that the instantaneous electrical current flowing in a particular external electrode due to the motion of the electron is

$$I_{ext}(t) = q\mathbf{E}_u \cdot \mathbf{v}(t), \tag{28.12}$$

where $\mathbf{v}(t)$ is the instantaneous velocity of the electron and $\mathbf{E}_u$ is the time-independent electric field at the location of the charge under the following circumstance: the moving electron is removed, the electrode of interest is raised to voltage $V = 1$ Volt, and all other electrodes are grounded. Since the voltage difference is 1 Volt, the unit of $\mathbf{E}_u$ is 1/distance, and $I_{ext}(t)$ is in Amps. What the theorem emphasizes is this: a current flows in the external electrodes when an electron is flowing outside the electrodes in the vacuum, *even when there is no physical injection or extraction of charge from the electrode at that instant*. This is the displacement current due to the change in electric field lines terminating at the electrodes.

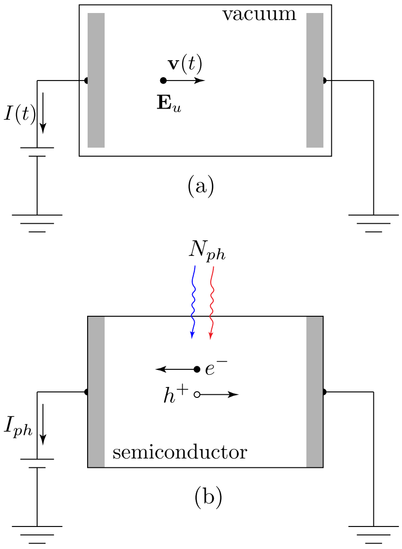


Fig. 28.13 (a) The Shockley–Ramo theorem, and (b) basic operation mode of semiconductor photodetectors. The primary task of a photodetector is to convert a photon flux N_{ph} into an electrical current I_{ph}.

Consider for example parallel-plate electrodes in the absence of the electron for which the electric field is $E = V/d$ where d is the distance between the electrodes. Then $E_u = 1/d$ at $V = 1$ Volt, and $I(t) = qv(t)/d$ at any voltage V, not just 1 Volt. One does not have to wait for the electron to reach the electrodes: the current changes instantaneously with the velocity of the electron. The electron may not travel in a vacuum: the same rule holds when it travels in a semiconductor. If photons of energy larger than the metal workfunction ($\hbar\omega > q\Phi_m$) were incident on the left electrode in Fig. 28.13 (a), electrons will be ejected, and their motion will create a current in the external circuit. This photoelectric effect is the fundamental principle of operation of most photodetectors.

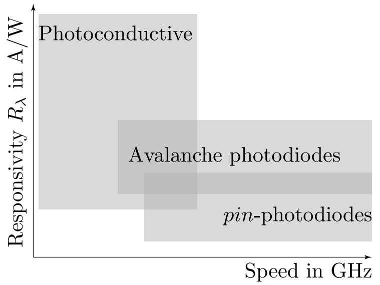


Fig. 28.14 Metrics of responsivity and speed of various photodetectors (qualitative illustration).

28.4 Photodetectors

If the vacuum is replaced by a semiconductor or semiconductor heterostructures (Fig. 28.13 (b)), an incident monochromatic light beam

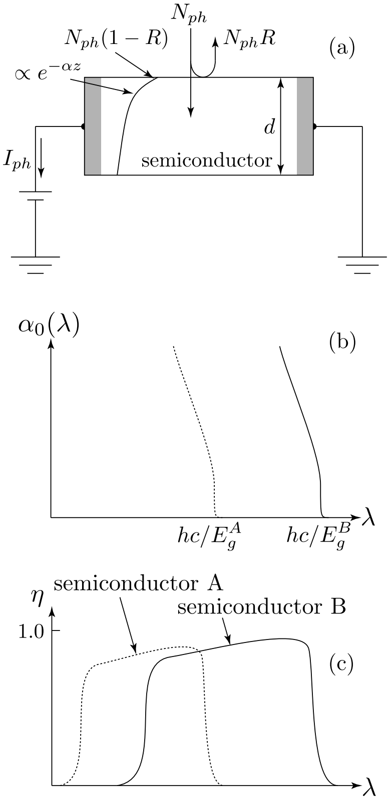


Fig. 28.15 (a) Metrics of a photodetector, (b) Schematic equilibrium absorption coefficients of two semiconductors as a function of the wavelength, and (c) Schematic quantum efficiencies of the photodetector.

of power P_{inc} produces a photon flux of $N_{ph} = P_{inc}/\hbar\omega$ photons per unit area per unit time. This creates electron-hole pairs throughout the semiconductor at wavelengths selected by the semiconductor bandgaps, which results in a photocurrent I_{ph}. Unlike solar cells where the primary aim is the conversion of energy, the aim in semiconductor photodetectors is the detection of photons with high efficiency, high resolution, and at high speed. The **quantum efficiency** η of a photodetector is a measure of the conversion of the number of photons incident to number of electrons that produce photocurrent in the external circuit. The **responsivity** R_λ of a photodetector is the "end result": the ratio of the current produced to incident light power, measured in A/W.

Fig. 28.15 shows how the two photodetector metrics are quantified. Of the incident photon flux, a fraction R is reflected, and only $1-R$ enters the semiconductor to produce electron-hole pairs. If d is the thickness of the photodetector and α is the absorption coefficient of the semiconductor for that photon energy, a fraction $(1-e^{-\alpha d})$ of the photons are absorbed in the semiconductor to create electron-hole pairs. Of these electron-hole pairs, a fraction η_c contributes to the electrical current flowing in the external circuit. Thus, the quantum efficiency and the responsivity of the photodetector are given by

$$\boxed{\eta = \frac{I_{ph}/q}{P_{inc}/\hbar\omega} = (1-R)\cdot(1-e^{-\alpha d})\cdot\eta_c,} \text{ and } \boxed{R_\lambda = \frac{I_{ph}}{P_{inc}} = \eta\cdot\frac{q\lambda}{hc}}. \tag{28.13}$$

The physical mechanism of the conversion process is $\hbar\omega \rightarrow e^- + h^+ \rightarrow I_{ph}$. The processes by which the photogenerated electron-hole pairs may fail to contribute to the external current are (a) radiative recombination before the electron-hole pair reach the electrodes, resulting in the re-emission of a photon, or (b) non-radiative recombination before the electron-hole pairs reach the electrodes that can occur in the bulk or near the surface which ultimately results in the emission of phonons. Therefore the design of a high efficiency photodetector is a game of maximizing η_c, and simultaneously minimizing R and $e^{-\alpha d}$. The reflectivity is minimized by the use of anti-reflection coatings, and α is maximized for the desired wavelength by choosing the semiconductor structure with a high absorption coefficient. The fraction η_c is maximized by the use of pure, defect-free semiconductors to prevent non-radiative recombination processes in the bulk, and especially by passivation to prevent non-radiative recombination near the surface where the absorption is the highest. The net quantum efficiency of semiconductors take the shape shown in Fig. 28.15 (c) due to the combined effects of all factors. The long-wavelength cutoff is for photons of energy smaller than the bandgap (the absorption coefficients are shown in Fig. 28.15 (b)), and the short-wavelength cutoff is due to surface losses.

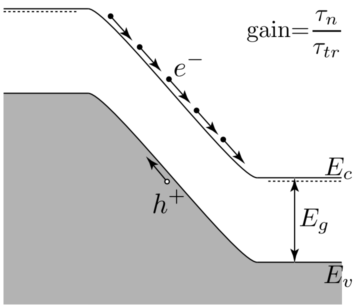


Fig. 28.16 For every hole created by the absorption of a photon, to maintain charge neutrality τ_n/τ_{tr} electrons need to be injected from the left electrode. This is the gain in photoconductive photodetectors.

There are several kinds of photodetectors, as indicated in Fig. 28.14. The simplest ones are **photoconductive** photodetectors, which are doped n- or p-type semiconductors with ohmic contacts. In a photoconductive detector, the base photocurrent is given by $I_{ph} = \eta q \frac{P_{inc}}{\hbar\omega}$ and the photocurrent in the presence of gain is $\Delta I = I_{ph} \cdot \frac{\tau_n}{\tau_{tr}}$, where η is the quantum efficiency. Here τ_n is the minority carrier lifetime in the p-type semiconductor, and τ_{tr} is the transit time of carriers between electrodes. This leads to a photoconductive responsivity

$$R_\lambda = \frac{\Delta I}{P_{inc}} = \frac{\eta q \frac{\tau_n}{\tau_{tr}}}{\hbar\omega} = \frac{\eta q \lambda}{hc} \cdot \frac{\tau_n}{\tau_{tr}} = \eta \cdot \frac{\lambda(\mu\text{m})}{1.24} \cdot \frac{\tau_n}{\tau_{tr}}, \tag{28.14}$$

which is in units of A/W, or amperes of photocurrent produced per watt of incident photon power. The ratio τ_n/τ_{tr} is similar to the ratio of minority carrier lifetime to the base transit time of a bipolar transistor that was discussed in Chapter 18. It was identified why this produces **gain**: the electrodes have to inject extra majority carriers to maintain charge neutrality in the semiconductor regions (Fig. 28.16). As a result, for $\eta \sim 1$, R_λ for photoconductive detectors can be as high as 100–1000 A/W in the 1–10 μm window. The price paid for this gain and the high responsivity is a low speed.

As opposed to photoconductive photodetectors, pn junction diodes have a built-in electric field at the junction. Since this region of high electric field is rather narrow in a pn junction, a thick i-layer is inserted to make a **p-i-n junction photodiode** (Fig. 28.17 (a, b)). A photovoltage is then produced in open-circuit situations, similar to the solar cell. Note the similarity of Fig. 28.17 (a) to Fig. 28.3 (a). The p-i-n photodiode can be operated both in the forward and reverse-bias modes for the purpose of photodetection: its $J - V$ characteristics are already calculated in Fig. 28.6 for the specific case when the incident light is the solar spectrum. When the *pin* photodetector is operated in the reverse-bias, a large electric field is experienced by the photogenerated carriers with a low current. The large field leads to a large carrier velocity and a fast response time, making high speed detection possible in such photodetectors. Note that the photogenerated carriers in the neutral p and n regions have low velocities and are prone to be lost to recombination, as opposed to the carriers generated in the i-regions which are immediately swept out into the electrodes.

The generic spectral **responsivity** of a semiconductor photodetector is shown in Fig. 28.17 (c). The long-wavelength cutoff is because the absorption coefficient is small for photon energies below the energy bandgap. The short-wavelength cutoff is because the absorption coefficient is very large for energies far larger than the bandgap, and most light is absorbed near the surface and does not contribute to the photocurrent or photovoltage. A standard *pin* photodiode has no gain of the same type as the photoconductive detector, and the responsivity is therefore $R_\lambda = \eta \cdot \frac{\lambda(\mu\text{m})}{1.24}$ A/W. Since η takes the shape shown in Fig. 28.15 (c), the resulting responsivity has the corresponding spectral window. The spectral window can be modified by the introduction of quantum wells or dots inside the active region. Photodetector design and usability is dominated by enhancing the signal to noise ratio, as discussed in Problem 28.8.

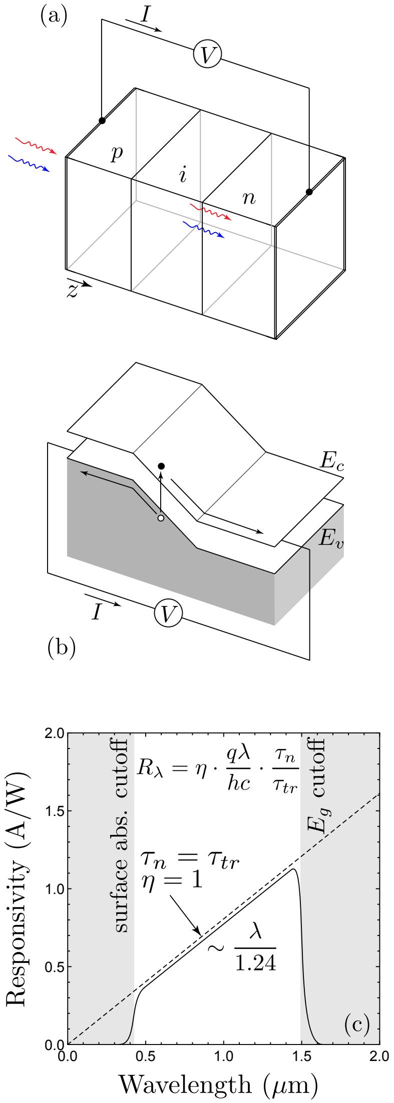


Fig. 28.17 (a) A p-i-n photodiode, (b) the energy band diagram at reverse-bias showing the electron and hole collection at electrodes, and (c) its responsivity as a function of the wavelength of the incident photons. The corresponding spectrum of InGaAs p-i-n diodes is between $1.0\mu\text{m} \leq \lambda \leq 1.7\mu\text{m}$, making it the preferred material for optical communication systems operating at 1.55μm.

28.5 Avalanche photodiodes

Fig. 28.18 shows the construction and operation of an avalanche photodiode, commonly known as APDs. The APD enhances the current per photon compared to a standard p-i-n diode by using avalanche multiplication process, which was discussed in Chapter 24. If the electron impact ionization coefficient is much larger than that of holes, the design uses a $n^+/p/p^-/p^+$ diode as shown in Fig. 28.18 (a). Under a large reverse-bias voltage, the depletion charge profile is shown in Fig. 28.18 (b), and the corresponding electric field and energy band diagram is shown in Fig. 28.18 (c, d). In a small thickness in the n^+/p junction on the p-side, the electric field exceeds a critical value to cause electron multiplication.

The energy band diagram in Fig. 28.18 (d) shows how a photon incident on the thick absorption layer creates an electron–hole pair. The hole moves towards the p^+ electrode and is collected there. As the electron traverses the steep conduction band-edge drop, it causes impact ionization. The net carrier multiplication factor M is indicated in Fig. 28.18 (e) as a function of voltage and temperature. Because at high temperatures the rate of optical phonon emission by hot electrons is enhanced, impact ionization is weakened, and the multiplication factor drops. Because of the carrier multiplication gain, APDs can be used in the **Geiger-mode** operation to detect single photons. Such devices, called single-photon avalanche diodes (SPADs) are important for quantum communications with single photons. The challenges in the use of APDs are the need to rapidly quench (or reset), and the noise due to the avalanche process.

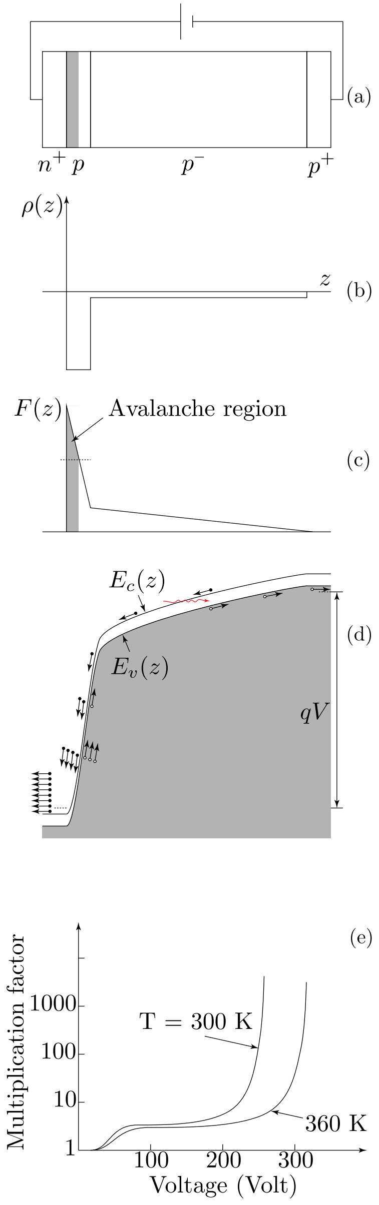


Fig. 28.18 Semiconductor avalanche photodiode (not exact, meant for illustration). (a) Layer structure, (b) charge profile, (c) electric field profile, (d) energy band diagram, and (e) carrier multiplication factor M vs. the applied reverse-bias voltage.

28.6 Quantum well infrared photodetectors

In Chapter 27 we discussed the absorption spectra $\alpha(\hbar\omega)$ of semiconductors for several physical mechanisms. The mechanisms were interband optical absorption, transitions between dopants and bands, and interband and intersubband absorption in quantum wells. Specificially, in Chapter 27, Section 27.12, we discussed the mechanism and the absorption coefficient when photons cause intersubband transitions in quantum wells. Fig. 28.19 shows how this transition is used for the detection of IR photons in quantum-well infrared photodetectors (QWIPs). The ground state is filled with electrons typically by doping, and the photons incident must have an electric field component perpendicular to the plane of the quantum well so that the transition has a non-zero matrix element. The electron excited to the higher energy state is thermionically emitted, or tunnels out in the presence of a DC electric field leading to a photocurrent. Typically a stack of several quantum wells are used, and QWIP arrays are used for the detection of $3 \leq \lambda \leq 20$ μm wavelengths which are accessed by the design of the quantum well eigenvalues. Arrays of QWIPs are used for a wide variety of applications ranging from thermal imaging for weather and geology to space missions. The number of photons that can be detected and distinguished from sources of noise depends on the mechanism of detection, and the dominant sources of noise in each type of photodetector (see Exercise 28.8). IR detectors of such kinds are of high interest for applications in night-vision, chemical spectroscopy, telecommunications and LIDAR.

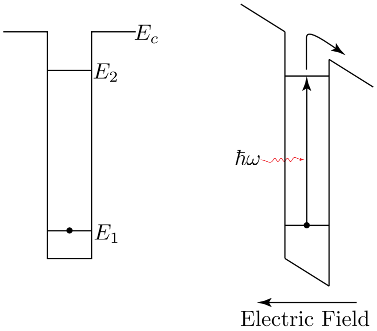


Fig. 28.19 Quantum Well Infrared Photodetectors.

28.7 Electro-absorption modulators

In addition to solar cells and photodetectors, the absorption of photons serve a very important role in modern communication systems. If the absorption coefficient of a semiconductor can be changed by an electric field by the amount $\Delta\alpha$, the net light absorbed by it changes by $e^{-(\Delta\alpha)L}$. In Chapter 27 Section 27.11, we discussed the Franz–Keldysh effect: an electric field causes a red-shift in the absorption spectrum of a bulk semiconductor. Electro-absorption modulators based on the Franz–Keldysh effect directly integrated with on-chip lasers are used today in communication systems. Such electro-absorption modulators have a modest $\Delta\alpha$, but a large L when integrated on a chip with waveguides.

A large $\Delta\alpha$ is obtained by using the similar effect in quantum wells. Fig. 28.20 shows an electro-absorption **modulator** that uses the quantum confined Stark effect. Fig. 28.20 (a) and (b) indicate the red-shift due to the quantum-confined Stark-effect in the absorption spectrum. A strong excitonic peak is observed in the absorption spectrum of quantum wells because the exciton binding energy is nearly 4× that of the bulk value and prominent at room temperature (see Table 27.2). The shift in the edge is substantial and is the basis for QCSE based electro-absorption modulators. Multiple quantum wells are inserted in the *i*-layer of a *pin* diode and electronic data as voltage pulses change a continuous stream of photons from a laser into the optical signal (Fig. 28.20 (c)) that is coupled into a waveguide for communication. Electro-absorption modulators today can transmit data at speeds at several 10′s of Gbps. The importance of modulators is increasing due to the interest in photonic integrated circuits.

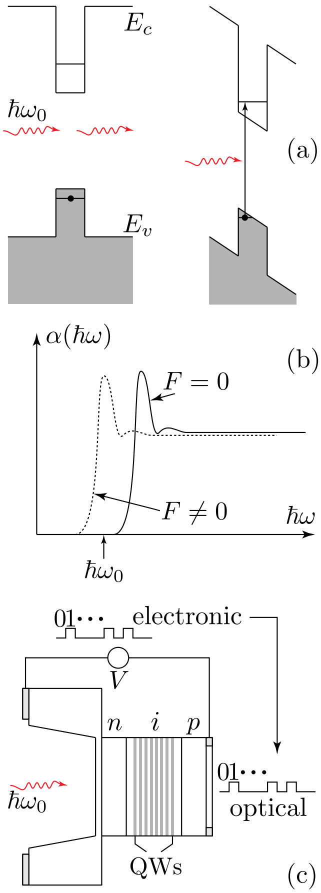


Fig. 28.20 Electro-absorption modulator based on the quantum confined Stark-effect.

28.8 Solar blind photodetectors

Photodetectors whose spectral absorption window fall outside of the solar spectrum are essential for imaging and detection of objects that can "hide in broad daylight". These photodetectors detect photons of UV wavelengths, typically shorter than 300 nm (4.13 eV). They are used for the detection of UV radiation, and in detecting and tracking missiles in the sunlight. Wide bandgap semiconductors such as Al(Ga)N are typically used for such photodetectors.

28.9 Chapter summary section

In this chapter, we learned:

- The principle of operation of solar cells, their quantum and thermodynamic efficiency limits, and the choices of semiconductor structures that maximize the solar energy conversion,
- The principle of operation of photodetectors and tradeoffs between photoconductive and pin detectors, their various uses, and the choices of semiconductors and mechanisms for various wavelength ranges of operation, and
- How electro-absorption modulators operate and how they are used for high-speed communications.

Further reading

Physics of Photonic Devices by S. L. Chuang devotes a chapter to photodetectors and solar cells. **Physics of Semiconductor Devices** by S. Sze has excellent chapters on photodetectors and solar cells, and **Semiconductor Optoelectronic Devices** by P. Bhattacharya has a nice balance between the underlying physics and several applications of the topics discussed in this chapter.

Exercises

(28.1) **Physics of semiconductor solar cell efficiencies**

(a) Verify Equation 28.2, and make the plot shown in Fig. 28.5.

(b) Calculate using a graphical method the efficiencies for solar cells made of semiconductors of bandgap 0.6 eV (e.g. Ge), 1.1 eV (e.g. Si), 1.4 eV (e. g. GaAs), and 3.4 eV (e.g. GaN).

(c) Using the equations derived for the current-voltage characteristics of solar cells, plot Fig. 28.6 for the bandgaps in part (b).

(d) Calculate and reproduce Fig. 28.7 for the efficiency vs. the bandgap for single-junction solar cells.

(e) Use the same graphical technique as in Fig. 28.10 to explore the efficiency limits of solar cells with higher number of junctions.

(28.2) **Photonic processes in semiconductor solar cells**

For red photons of energy $\hbar\omega \sim 1.8$ eV, the equilibrium optical absorption coefficient of Si with $E_g = 1.1$ eV is $\alpha_0(\hbar\omega) \sim 3 \times 10^3$/cm whereas that of GaAs with $E_g = 1.4$ eV is $\alpha_0(\hbar\omega) \sim 2 \times 10^4$/cm.

(a) Explain why it is so much higher for GaAs with representative sketches.

(b) How thick should a Si or GaAs layer be to absorb $(1 - e^{-3}) \sim 1 - \frac{1}{20} = 85\%$ of the photons that enter the semiconductor?

(c) The absorption of a photon creates an electron in the conduction band, and a hole in the valence band. Sketch this process using the bandstructure $E(k)$ first for GaAs and then for Si.

(d) In GaAs and Si, the extra electron in the conduction band and hole in the valence band must lose their excess energies and relax to the edges of the respective bands. What process enables this energy loss?

(e) Once they have reached the band edges, what process will get rid of the extra electron and hole? How is this process fundamentally different between GaAs and Si, and which is much faster? Why?

(f) In a solar cell, a p-n junction is used to separate the electrons and holes spatially and generate an open-circuit voltage well before they meet their undesired fates of part (e). Argue why the voltage generated will be close to the bandgap of the semiconductor.

(28.3) **Approximate results of solar cell characteristics**
Show that the approximate results of Equation 28.11 reproduce the exact results of Equation 28.8, except for by reproducing Figures 28.6 and 28.7 with the approximate results. Note the results differ for $E_g \lesssim k_b T_1$ and explain the reason why the approximation does not work for this case.

(28.4) **Solar cell efficiencies: The missing 20%**
Carefully read the two papers: 1) W. Shockley and H. J. Queisser, J. Appl. Phys., **32**, 510 (1961), and 2) C. H. Henry, J. Appl. Phys., **51**, 4495 (1980). Discuss in this context, the limits of efficiencies of single-junction, and multijunction solar cells. Answer the following:

(a) Show that the entropy of blackbody radiation of energy E from a source at temperature $T_\odot$ is $S = \frac{4}{3}\frac{E}{T_s}$.

(b) Show that the maximum conversion efficiency of solar energy is then $\eta_{\max} = 1 - \frac{4}{3}\frac{T_1}{T_\odot} \sim 93\%$ for $T_\odot = 5800$ K and $T_1 = 300$ K.

(c) Is it beneficial to cool down a solar cell to increase its efficiency?

(d) Using C. H. Henry's paper, describe each portion of the lost ~ 20 % that explains why the maximum multijunction cell efficiency is 72%.

(28.5) **Present and future of solar cells**
Write a report on the past, present, and future of solar cell energy generation and its role in the energy economy of the world. In particular, include the following:

(a) Why the highest efficiency multijunction cells are expensive to produce, and the cost-efficiency tradeoff,

(b) Which new materials can be expected to see increased use in the future,

(c) How much of the increasing energy need can truly be sustained by solar power plants on Earth, and

(d) What methods from a policy standpoint can enable a balanced approach towards harnessing more solar energy in an equitable manner in the future.

(28.6) **When heat and light compete: night vision**
Narrow bandgap semiconductors are used for night-vision cameras and heat mapping. Thermal imaging requires photodetectors that can absorb long-wavelength photons of small energies. Answer the following questions about the choice of semiconductor materials and challenges in making such photodetectors.

(a) What intrinsic problem is encountered if to detect photons of energy $\hbar\omega \leq k_b T$ by interband optical transitions? Explain the need to cool such long-wavelength photodetectors. Which semiconductors are typically used for such photodetectors?

(b) For special applications, materials other than semiconductors are used that convert the photon energy directly into heat, unlike a pulse of current or voltage in semiconductor photodetectors. Compare the advantages and disadvantages of this mode of long-wavelength photodetection.

(28.7) **SACM SWIR avalanche photodiodes**
Fig. 28.21 schematically shows the mechanism of operation of InGaAs p-i-n diodes used for detectors in the short-wave IR (SWIR) regime of the electromagnetic spectrum. The particular device is called the **S**eparate **A**bsorption, **C**harge, and **M**ultiplication Avalanche Photo Diodes (SACM APDs). Answer the following questions.

(a) The mechanisms of absorption, tunneling, and avalanche multiplication are indicated schematically in the energy band diagram. Using this energy band diagram and the field profile shown, explain how the device acts as an efficient photodetector.

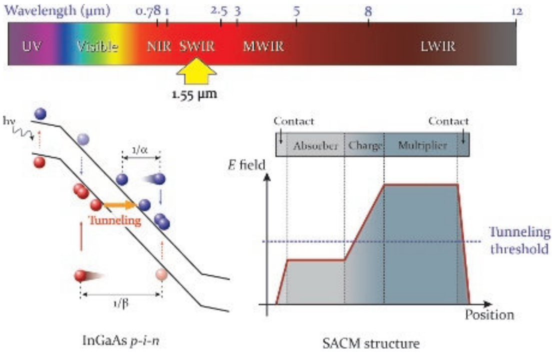


Fig. 28.21 Design of a SACM APD for the SWIR spectral window. Kindly shared by Prof. S. Krishna and H. Jung of the Ohio State University.

(b) Find the composition of In needed for the InGaAs for interband absorption edge at the optical communication wavelength $\lambda = 1.55$ μm.

(c) Using the impact ionization coefficients discussed in Chapter 24, Table 24.1, estimate the typical thickness of the multiplier region necessary for efficient APD operation.

(d) Do research on where such SACM APDs are used today and project their need in the future.

(28.8) Noise and frequency response of photodetectors

Using the Equation 28.13 for the photocurrent I_{ph}, in this problem the derivation of the frequency response of a photodetector is outlined.

(a) Show that the time-dependent response to a sinusoidal input power $P(t) = P_{inc}(1 + m\cos\omega t)$ where m is a dimensionless modulation factor, the photocurrent response is

$$I(t) = I_{ph}[1 + \frac{m}{\sqrt{1+(\omega\tau_n)^2}}\cos(\omega t - \phi)] \implies$$

$$\text{rms photocurrent: } i_p = q\eta\frac{mP_{inc}}{\sqrt{2}\hbar\omega}\frac{\tau_n}{\tau_{tr}}\frac{1}{\sqrt{1+(\omega\tau_n)^2}} \tag{28.15}$$

This is the desired signal.

(b) The following are the undesired noise in the current. Explain the meaning of each of the three noise terms:

$$\text{Shot noise: } \langle i_s^2(\omega)\rangle = 2q\langle I\rangle\Delta\omega$$

$$\text{Recomb. Gen. noise: } \langle i_{rg}^2(\omega)\rangle = \frac{4qI_{ph}\frac{\tau_n}{\tau_{tr}}\Delta\omega}{1+(\omega\tau_n)^2}$$

$$\text{Thermal noise: } \langle i_{th}^2\rangle = \int_0^{\Delta\omega}\frac{4}{R}\frac{\hbar\omega}{e^{\frac{\hbar\omega}{k_bT}}-1}d\omega \approx \frac{4k_bT}{R}\Delta\omega, \tag{28.16}$$

where $\Delta\omega$ is the bandwidth in 1/s, R is the resistance of the photodetector, and T is the temperature that cause blackbody radiation inside the photodetector.

(c) The ratio of their powers is the SNR:

$$SNR_{power} = \frac{i_p^2}{\langle i_{th}^2\rangle + \langle i_{rg}^2(\omega)\rangle}$$

$$= \frac{\eta m^2\frac{P_{inc}}{\hbar\omega}}{8\Delta\omega[1+\frac{k_bT}{qRI_{ph}}\frac{\tau_n}{\tau_{tr}}(1+(\omega\tau_n)^2)]}. \tag{28.17}$$

Where the last equation gives the net signal to noise ratio in terms of all known quantities. Find typical quantitative values of the SNR at room temperature, and discuss how the SNR changes with temperature.

(28.9) The quantum mechanic

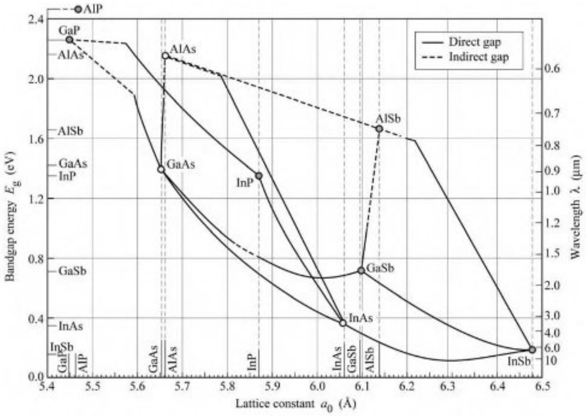


Fig. 28.22 Bandgaps and lattice constants of compound semiconductors (from "Light Emitting Diodes" by E. F. Schubert).

As the CTO of the stealth startup "The Quantum Mechanics" you win a project from NASA to design an infrared photodetector for the detection of a very weak $\lambda_0 = 12.4$ μm radiation from a distant star. Because you have in your team the world's best group of III-V compound semiconductor MBE growers who can grow for you layered heterostructures of any combination of materials you can dream up from Fig. 28.22, you decide the design will be done using *intersubband transitions* in III-V quantum wells. Describe your design in quantitative detail – including plots, choices of materials, thicknesses, etc. Pay attention to the *selection rules* for optical transitions, assume the 75%:25% $\Delta E_c : \Delta E_v$ band offset distribution, and use the rule of thumb that you cannot grow a strained layer thicker than $t_{cr} \sim \frac{a_0}{2\epsilon}$ where ϵ is the strain. Feel free to use the results of Exercise 14.2. Explain if your IR photodetector will work at room temperature, or whether it needs to be cooled. Cooling may not be a problem, especially if it goes to space!

29

Reach for the Stars: Lasers and LEDs

29.1 Lasers and LEDs: a résumé 773
29.2 Einstein's A and B coefficients 776
29.3 Semiconductor emission 778
29.4 Entropy of optical transitions 779
29.5 Gain and emission in bands 781
29.6 Spontaneous emission: LEDs 785
29.7 Stimulated emission: lasers 787
29.8 Double heterostructure lasers 789
29.9 Laser rate equations 792
29.10 Case study: blue laser diode 795
29.11 DFB, VCSELs and QCLs 796
29.12 Towards field quantization 798
29.13 Broadband field modes 803
29.14 Quantization of fields 807
29.15 Field quantization: aftermath 810
29.16 Quantized light-matter interaction 814
29.17 Fundamental optical processes 819
29.18 Einstein's re-derivation of Planck's law: back to 1^2! 821
29.19 Chapter summary section 823
Further reading 824
Exercises 825

In this penultimate chapter of the book, we discuss the most revolutionary photonic device of the last century, the laser. We discuss how the idea of a laser can be concluded from the light-matter interaction principles provided by Einstein. The laser took several decades to realize, in spite of all ingredients being available. We discuss how semiconductors and their heterostructures are used to make the lasers that power today's information age by moving massive amounts of data at the speed of light. In this process, we also learn about the operation of light-emitting diodes (LEDs), and connect to advanced topic of field-quantization. The goals of this chapter are to learn:

- How does the absorption coefficient of semiconductors change under non-equilibrium conditions?
- What are the requirements for spontaneous and stimulated emission of light from semiconductors?
- How do LEDs based on spontaneous emission operate and what are their quantum limits of efficiency?
- What are the principles of operation of semiconductor lasers, and how is stimulated emission sustained?
- How can we reconcile the quantum nature of light by quantizing the electromagnetic field?

29.1 Lasers and LEDs: a résumé

The first *optical* laser was realized in 1960 (see Fig. 29.1), a few years after the demonstration of its lower frequency microwave counterpart – the maser (microwave amplification by stimulated emission of radiation). The optical laser is conceptually identical to its microwave counterpart, but its oscillation frequency is $\sim 1000\times$ higher. The laser produces *coherent and monochromatic* light with a very small spread in wavelengths (or frequency). The history of the development of the maser and the laser form an interesting case study. Unlike several earlier scientific discoveries where experiments predated theory and understanding, the basic idea of the laser and the necessary ingredients needed to realize it were available for roughly three decades before its successful realization. Why did it take so long?

Fig. 29.1 Theodore Maiman realized the first optical laser in 1960. Seen here next to the ruby laser that emits coherent light at $\sim$694 nm in kW pulses.

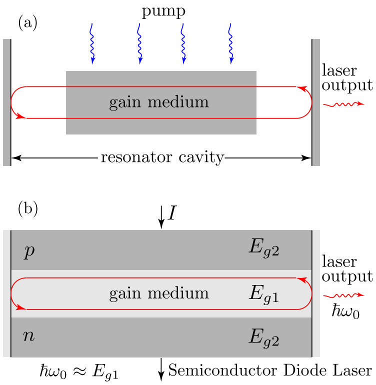


Fig. 29.2 (a) The building blocks of a laser. A gain medium is produced by raising matter into an excited state with a pump. Photons emitted into modes allowed by the resonator cavity cause stimulated emission into the same mode. A fraction of the coherent and monochromatic photons are coupled out as the laser output. (b) A semiconductor laser.

Fig. 29.3 Charles Townes, the inventor of the maser for which he was awarded the Nobel Prize in Physics in 1964 with Nikolay Basov and Alexander Prokhorov. In an article on the development of lasers, Townes recounts "*... My discussion with Von Neumann had a more special twist and occurred at a social occasion. When I told him about our maser oscillator, he doubted the uncertainty principle allowed our observation of such a narrow frequency width to be real. After disappearing in the crowd for about 15 minutes, he came back to tell me he now understood the situation; my argument was correct. That Von Neumann took even that long to understand impressed me. He then went on to urge that I try for stimulated emission effects in the infrared region by exciting electrons in semiconductors. ...*". From IEEE J. Quantum Electronics, 20, 547 (1984) an issue on lasers (recommended reading).

The key idea that makes a laser possible is traced back to 1917, when Einstein introduced the concepts of **spontaneous and stimulated emission** in his re-formulation of Planck's blackbody radiation formula. He did this to clarify the role of "matter" in light-matter interaction (we discuss his reformulation in Sections 29.2 and 29.18). Quantum theory was formulated in the next decade, and by 1930, the necessary ingredients to understand and compute energies of atoms, molecules, and solids were in place. Energies of atoms and molecules could be computed and agreed with experiment with spectacular accuracy. These are the ingredients needed for a laser.

Reduced to its essence, a laser (Fig. 29.2) consists of *collection* of matter (several atoms or molecules in a gas, liquid, or a solid) that is raised to an excited state, such that it is "primed" to relax to its ground state by the emission of photons of a **very specific wavelength**. Ordinarily, the atoms act independently: they relax whenever they like (i.e., incoherently). The photons they emit therefore go in all directions with no specific phase differences between them. This is **spontaneous emission**. But if by some means the emission is severely restricted to only one (or a few) modes by placing a resonant cavity around the excited matter, the passage of a photon of that mode will stimulate the atoms to relax by emitting the photon exactly into that specific mode. This **positive feedback** forces the collective and coherent relaxation of the excited matter. They then emit coherent photons into a mode of a specific wavelength. A photon in a particular mode produces clones of it as it bounces back and forth between the walls of the resonant cavity (or mirrors), thus building up a beam of coherent photons, all of a very precise wavelength. This is **stimulated emission**, which is a means of self-replication or **optical gain**.

When the gain exceeds the losses in the medium, a laser is realized. The difference between spontaneous and stimulated emission in the sharpness of wavelengths of photons emitted is rather dramatic: for example in semiconductor lasers it can reduce from 100 Å to well below 0.1 Å, with a corresponding dramatic reduction in the uncertainty in frequency $\delta\omega$. Thus a laser is a **very precise oscillator**. In fact its precision is so high that it seemed to the founders of quantum mechanics to violate the uncertainty principle! The reason why in spite of the conditions being ripe for the invention of the laser (or the maser) in the 1930s, it still took so long to actually realize it is discussed eloquently in articles by Townes (Fig. 29.3) and Schawlow (Fig. 29.4), two of the pioneers of this device. Whereas the concepts of positive feedback and resulting sharp oscillation frequencies obtained in cavity resonators was obvious to electrical engineers who designed microwave circuits during and after World War II, they were not that commonly known to physicists who studied atomic spectra.

At first glance, a laser *seemingly* violates the foundational rules of quantum mechanics and thermodynamics. The first *apparent* violation is mentioned above due to its very narrow frequency spread is the energy-time uncertainty principle. The second is of thermal equilibrium. The laser is possible only when matter is taken far out of equilibrium such that the occupation of excited states far exceeds that of the ground state – this is called **population inversion**. The equilibrium occupation is the exact opposite: an inverted population follows the statistics of **negative temperature**!

The laser, since its birth till today, has continued to refine and test the core concepts of quantum mechanics and thermodynamics. The first laser was realized by Townes in 1954 in the microwave frequency using the transitions of ammonia molecules. Townes and Schawlow in 1958 outlined how one could obtain an optical frequency laser with a gain medium realized by stimulated emission, combined with a resonator cavity made of mirrors. Their proposal set off several experimental efforts. The first optical laser was realized by Maiman in 1960 using transitions of Cr atoms embedded in sapphire (Al_2O_3), or what is known as a ruby crystal. The crystal was shaped into a rod, and its ends were polished into mirrors to form the resonator. This first ruby laser produced high-power ($\sim$kW) pulses of coherent light. Soon thereafter, the first **continuous-wave** (or CW) laser was realized by Ali Javan using transitions of He-Ne gas mixture. It was realized that the transitions need not be that of electronic states alone, it could also be the **vibrational modes** of molecules. An important laser of this kind is the CO_2 laser invented by C. K. N. Patel. This laser emits at 10s of μm wavelength, has overall electrical to optical energy conversion efficiency in excess of 10%, and can produce enormous output powers sufficient to cut steel and weld metals. Methods of compressing the light pulse using Q-switching and mode-locking can now produce laser pulses well below femtosecond time scales and GW energies to probe the extremes of time and energy. The shortest laser pulses today can reach **nearly one cycle of oscillation**!

Fig. 29.4 Art Schawlow, one of the pioneers of the laser. He was awarded the Nobel prize in physics in 1981. He humorously lobbied for the use of the correct name "Light-Oscillator by Stimulated Emission of Radiation" over "LASER", but it is not difficult to see why it did not catch on.

Fig. 29.5 Nick Holonyak, a semiconductor pioneer, demonstrated the semiconductor light-emitting diode, and one of the first pn junction as well as quantum well lasers.

Semiconductor lasers "grew up" alongside solid-state and gas lasers. They are called **diode lasers** to distinguish them from other types of lasers, and more as an acknowledgement that their near-magical properties that put them in a very separate class from other lasers stem from the fact that at its heart is a **semiconductor pn junction**. Laser oscillation in semiconductor pn junctions was first reported nearly simultaneously by four groups (at GE Schenectady, GE Syracuse, IBM, and Lincoln Labs) in 1962. These observations were at low temperature, and under high electrical injection currents. The condition necessary for obtaining lasing oscillation in semiconductors was outlined in an influential 1962 paper by Bernard and Duraffourg. In 1963, Herbert Kroemer (Fig. 15.21) proposed the use of **double heterostructures** to significantly lower the lasing threshold for semiconductors. By 1970, Hayashi and Panish, and Alferov(Fig. 29.39)'s group had obtained room-temperature lasers using the double heterostructure idea. Dyment offered an explanation of the far-field radiation patterns and the typical multi-mode lasing operation of semiconductor lasers. Kogelnik and Shank in 1970 introduced distributed feedback Bragg (DFB) mirrors that enabled single-mode oscillation in semiconductor lasers. Significant reduction of lasing thresholds were achieved in the 1980s–2000s by introducing quantum wells and quantum dots, and with clever use of strain in the active regions. The laser geometry itself has gone through several refinements: variants such as edge-emitting DFB lasers and vertical cavity surface emitting lasers (VCSELs) are currently in production. The spontaneous emission regime of operation is exploited in semiconductor pn junctions for LEDs, which do not require feedback and resonators.

As opposed to the cm to meter-size optically pumped solid-state and gas lasers, a semiconductor laser is **electrically pumped**, less than 1 mm^3 in size and delivers watts of power. Their efficiency of energy conversion can exceed 50%, their lifetimes are measured in 100s of years. Importantly, they can be **modulated** at 10s of GHz speeds by directly modulating the injected current which creates the population inversion in the first place. All of these can be achieved in an integrated chip, which makes them marvels of modern technology, and powers high speed data transfer of enormous bandwidths. Mixing the lasers also pays rich dividends: for example, diode-pumped solid-state (DPSS) lasers combine the best of two technologies. The uses of lasers extend beyond communication systems into numerous areas that range from astrophysics (e.g. gravitational wave detectors) to industrial cutting and welding, laser eye surgery, lighting, and beyond.

In this chapter, we focus on the fundamental physics of the laser. We build on the concept of the optical absorption coefficient of semiconductors introduced in Chapter 27 to first explain spontaneous and stimulated emission. We then show how this leads naturally to the many avatars of a semiconductor p-n diode as a light emitter: as a light-emitting diode (LED) that uses spontaneous emission, to semiconductor optical amplifier (SOA) with population inversion, and as a laser upon the addition of resonators. Instead of describing various kinds of lasers for which there are excellent references mentioned in the further reading section, we choose to do a case-study of a specific blue laser diode. Then we dive deeper into a discussion of the difference between radiation from a laser and blackbody radiation to introduce field quantization in the last part of this chapter.

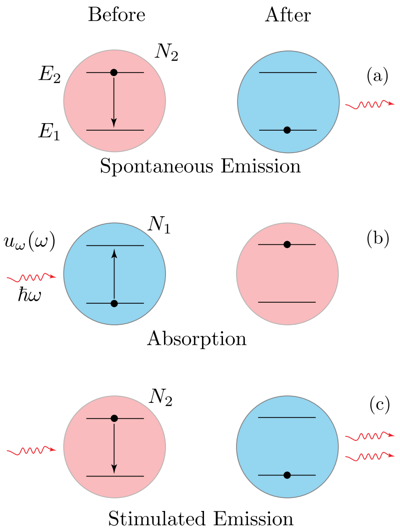


Fig. 29.6 Einstein's introduction of stimulated emission as one of the three fundamental optical processes.

29.2 Einstein's A and B coefficients

The conceptual breakthrough that led to the laser is traced back to Einstein's uncovering of the mechanism of **stimulated emission** in 1917 (in A. Einstein, "On the Quantum Theory of Radiation" Z. Phys. 18, 121 (1917).). Up until that time, emission of radiation from excited atoms was thought to be only due to the spontaneous process (Fig. 29.6 (a)). The rate of spontaneous emission does not depend on the prior presence of photons. The rate of absorption (Fig. 29.6 (b)) of course depends on the availability of photons characterized by a spectral density $u_\omega(\omega)$. Einstein realized that if these were the only two processes, it is impossible for a collection of atoms to come into thermal equilibrium with a blackbody radiation source[1]. He deduced that a third process: **stimulated emission** is necessary to attain equilibrium. Stimulated emission is depicted in Fig. 29.6 (c): the new photon emitted is identical in every way to the one that stimulates the process. The end result of Einstein's theory is that the **absorption rate is proportional to N_{ph}, but the emission rate is proportional to $1 + N_{ph}$ where the 1 is due to spontaneous emission and the proportionality to N_{ph} is due to stimulated emission**. Einstein's classical thermodynamic result is found to hold in a full quantum treatment of Section 29.18, and underpins two semiconductor photonic devices that emit light: the LED and the laser. We begin our analysis with Einstein's semi-classical derivation, leaving the full quantum treatment to the last sections of this chapter.

Consider N_1 atoms in which the electron is in the ground energy state E_1, and N_2 atoms for which the electron is in the excited energy state E_2, as indicated in Fig. 29.6. Let the rate of spontaneous emission of photons be $A_{21} \cdot N_2$ – it is proportional to the number of atoms that can emit, and is independent of the number of photons present. The rate of absorption is $B_{12} N_1 L(\omega)$, which is proportional to the photon density $L(\omega)$ and the number of atoms that can absorb. Then, the rate of stimulated emission must be $B_{21} N_2 L(\omega)$, which is proportional to the number of atoms that can emit, and the number of photons present. Here A_{21}, B_{21}, and B_{12} are the Einstein A and B coefficients. **At thermal equilibrium there must be detailed balance between total absorption and emission**, *and* the resulting photon density $L(\omega)$ in the electromagnetic field must be the Planck distribution[2]:

$$A_{21} N_2 + B_{21} N_2 L(\omega) = B_{12} N_1 L(\omega)$$

$$\implies L(\omega) = \frac{\frac{A_{21}}{B_{21}}}{\frac{B_{12}}{B_{21}} \frac{N_1}{N_2} - 1} = \frac{\frac{\omega^2}{\pi^2 \hbar c^3}}{e^{\frac{\hbar\omega}{k_b T}} - 1}$$

$$\frac{N_1}{N_2} = e^{\frac{\hbar\omega}{k_b T}} \implies B_{12} = B_{21}, \text{ and } \frac{A_{21}}{B_{21}} = \frac{\omega^2}{\pi^2 \hbar c^3}. \tag{29.1}$$

If one of the coefficients (say the absorption coefficient $B_{12} = B$) is known, then all coefficients are known[3]. Importantly, **these coefficients remain the same in non-equilibrium** and thus provide the rates and spectra of photon emission and absorption. The rates in thermal equilibrium is then, using $B = B_{21} = B_{12}$ and $A = A_{21}$

$$R_{abs} = B N_1 L(\omega), \; R_{stim} = B N_2 L(\omega) \text{ and } R_{spon} = A N_2$$

$$R_{em} = A N_2 + B N_2 L(\omega) = A N_2 (1 + \frac{B}{A} L(\omega)) = A N_2 (1 + N_{ph}), \tag{29.2}$$

where it is seen that the emission rate is split into the spontaneous part proportional to 1 and the stimulated part proportional to N_{ph}, the number of photons. $R_{abs} = R_{em}$ at thermal equilibrium. The stimulated emission process in Fig. 29.6 (c) vividly illustrates that it is a process of **optical gain or amplification**: the number of photons exiting is larger than those entering the medium. The stimulated emission process can dominate over absorption when $N_2 > N_1$. This condition of population inversion is impossible under thermal equilibrium since it would correspond to a negative temperature. Thus the optical gain, a key ingredient for a laser, is only available under non-equilibrium conditions. Fig. 29.7 is an illustration of the mechanism of optical gain. It highlights that for sustained gain, the atomic system must be continuously pumped and the condition $N_2 > N_1$ must be maintained. In non-equilibrium conditions the A and B coefficients stay the same as Equation 29.1, but the rates become $R'_{abs} = BN'_1 L'(\omega)$, $R'_{stim} = BN'_2 L'(\omega)$, and $R'_{spon} = AN'_2$, where the primed symbols indicate non-equilibrium values, and emission can exceed absorption.

---

[1] Recall from earlier chapters that the blackbody radiation energy per unit volume per mode interval $(\omega, \omega + d\omega)$ is the product of the photon density of states $\frac{\omega^2}{\pi^2 c^3}$, the photon energy $\hbar\omega$, and the photon number $N_{ph} = \frac{1}{e^{\frac{\hbar\omega}{k_b T}} - 1}$ in the expression $u_\omega(\omega) = \frac{\omega^2}{\pi^2 c^3} \cdot \hbar\omega \cdot N_{ph}(\omega) = \frac{\hbar\omega^3}{\pi^2 c^3} \frac{1}{e^{\frac{\hbar\omega}{k_b T}} - 1}$. The **photon density** per unit energy interval per unit volume $L(\omega)$ is given by the relation $\hbar\omega \cdot L(\omega) \cdot d(\hbar\omega) = u_\omega(\omega) d\omega \implies L(\omega) = \frac{\omega^2}{\pi^2 \hbar c^3} \cdot \frac{1}{e^{\frac{\hbar\omega}{k_b T}} - 1}$ in $\frac{1}{\text{cm}^3 \cdot \text{eV}}$ units.

[2] Note that $E_2 - E_1 = \hbar\omega$, and the ratio $N_1/N_2 = e^{\frac{\hbar\omega}{k_b T}}$ at thermodynamic equilibrium. For semiconductors, the discrete levels E_2 and E_1 are replaced by band states in the conduction and valence bands respectively.

[3] Chapter 27 was dedicated to the evaluation of the absorption coefficients $\alpha(\hbar\omega)$ in semiconductors. The Einstein B coefficient is directly related to this absorption coefficient, as will be used shortly.

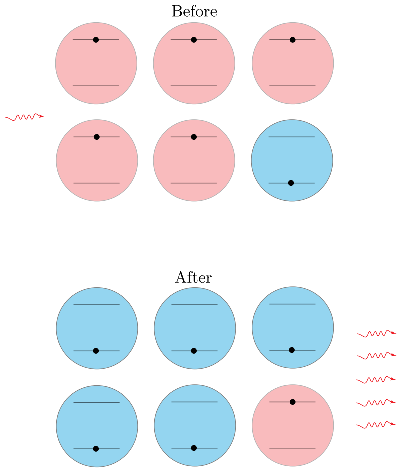


Fig. 29.7 A possible scenario illustrating the origin of optical gain in an atomic laser system. In such lasers, the intensity of light I which is proportional to the number of photons changes with distance z as $\frac{dI}{dz} = \sigma(N_2 - N_1) \cdot I = gI$, where σ is the gain cross-section (in cm^2), and N_2 and N_1 are the volume densities (in cm^{-3}) of atoms in the excited state and ground state respectively. The gain coefficient $g = \sigma(N_2 - N_1)$ has units of cm^{-1}, the same as the absorption coefficient, and the intensity changes as $I(z) = I(0) \cdot e^{gz}$.

29.3 Semiconductor emission

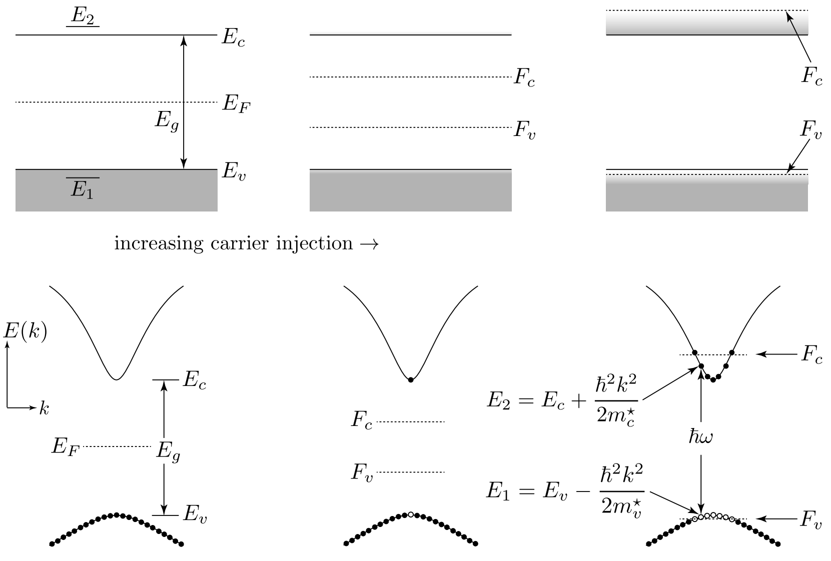
Einstein's ideas for atomic systems carries over to bulk semiconductors and quantized structures. Applying Einstein's theory for atomic systems to optical transitions in semiconductors requires three conceptual changes:

(1) Discrete energy levels → Density of states.
(2) Boltzmann statistics → Fermi–Dirac statistics.
(3) Conservation of crystal momentum **k** upon optical transition.

The reasons for (1) and (3) have been discussed at length in prior chapters. Because the atoms of an atomic laser system are far apart, the electron wavefunctions of adjacent atoms do not overlap. The electrons are independent and are bound to the same nucleus even when the atom is in the excited state. For such a dilute collection of atoms, the Boltzmann statistics is sufficient. But in a semiconductor the atoms are packed close, and electron wavefunction overlaps result in the broadening of discrete energies into bands (Fig. 29.8). The electron Bloch eigenfunctions are delocalized over the entire crystal – they do not belong to any particular atom. Because a large number of electrons can occupy an allowed energy in a semiconductor, there is no escaping the Pauli exclusion principle and the resulting Fermi–Dirac statistics discussed in Chapter 4.

We will shortly evaluate the consequences of these changes compared to the atomic system, but they may be anticipated from qualitative considerations. Because of the sharp discrete energies of atomic systems, they emit photons at sharp energies, and the spectrum of an atomic laser is very sharp. But because of the density of states and spread in energies in bulk and quantum well semiconductor structures, the emission spectra of a typical semiconductor laser is broader than an atomic laser[4]. On the other hand the closeness of the atoms in a semiconductor results in a very high density of states and high optical gain. Thus, semiconductor lasers can be made at far smaller

[4] Semiconductor quantum dot lasers allow sharp output spectra, but making high density of quantum dots of uniform size and emission wavelength remains a challenge in crystal growth.


E_2
E_c
E_F
E_g
E_v
E_1
F_c
F_v
increasing carrier injection →
$E(k)$
k
$E_2 = E_c + \frac{\hbar^2 k^2}{2m_c^\star}$
$\hbar\omega$
$E_1 = E_v - \frac{\hbar^2 k^2}{2m_v^\star}$

(a) Thermal Equilibrium

(b) Spontaneous Emission

(c) Stimulated Emission

Fig. 29.8 Semiconductor population inversion

size (~ 200 μm) compared to atomic lasers ($\sim 10 - 100$ cms). Semiconductor lasers can be modulated directly by current injection at very high speeds.

29.4 Entropy of optical transitions

Additional insight of the flow of energy in optical transitions is obtained from the analysis of entropy of the combined light-matter system. The advantage of an entropy point of view is that the results that are obtained are of complete generality, and will apply no matter if the optical transitions are due to interband processes, or due to excitons, phonons, or dopants. Consider a scenario of a semiconductor layer of thickness t held in non-equilibrium (Fig. 29.8 (b) and (c)) such that the electrons in the conduction band have a quasi-Fermi level F_c and the electrons in the valence band have a quasi-Fermi level F_v, and $F_c - F_v = qV$. For example, this is the condition in the active region in the pn junction in a semiconductor LED or laser diode. Under these conditions, the absorption coefficient of this region of the semiconductor for upward transitions only[5] is $\alpha_{up}(\hbar\omega)$. When a beam of photons of frequency ω and occupation number n_ω passes through this region, the occupation number in the mode ω will change according to

$$\Delta n_\omega = -\underbrace{\alpha_{up}(\hbar\omega) \cdot t \cdot n_\omega}_{\text{abs.}} + \underbrace{a + b \cdot n_\omega}_{\text{spon. + stim. em.}}, \tag{29.3}$$

where the effective Einstein-like a coefficient is a measure of the spontaneous emission, and the b the coefficient of stimulated emission. At

[5] Here $\alpha_{up}(\hbar\omega)$ in cm^{-1} units is the absorption coefficient for *only upward transitions*, and hence the subscript. The downward stimulated transitions characterized by the coefficient b are excluded in this definition. Note that lowercase a and b are used to distinguish these dimensionless effective Einstein coefficients from those of other sections.

thermal equilibrium the net absorption is balanced by net emission. then, $\Delta n_\omega = 0$, yielding the steady-state occupation

$$\bar{n}_\omega = \frac{\frac{a}{b}}{\frac{\alpha_{up}(\hbar\omega)t}{b} - 1}. \tag{29.4}$$

Thermal equilibrium between this semiconductor and the photon field is precisely defined by the requirement that the entropy of the combined system is at a maximum[6]. In that case, consider the event of the emission of one photon of energy $\hbar\omega$ from the semiconductor when an electron drops from the conduction band to an empty state in the valence band. The semiconductor lowers its energy $U_i - U_f$ by $\hbar\omega$. But the *free energy* of the semiconductor is lowered by $F_i - F_f = F_c - F_v = qV$ because the electron was injected from a battery which held the quasi-Fermi levels $F_c - F_v = qV$ apart. The first law of thermodynamics requires $F = U - TS$ where T is the temperature and S the entropy, leading to a change in entropy of the semiconductor

$$F_i - F_f = (U_i - U_f) - T(S_i - S_f) \implies S_i - S_f = \Delta S_{\text{semi}} = \frac{qV - \hbar\omega}{T}. \tag{29.5}$$

Now the total change in the entropy of the photon (boson) field is given by considering only mode ω whose occupation increased exactly by $\Delta n_\omega = +1$:

$$S_{\text{photon}} = k_b \sum_\omega [(n_\omega + 1)\ln(n_\omega + 1) - n_\omega \ln(n_\omega)]$$

$$\Delta S_{\text{photon}} = k_b \sum_\omega \Delta n_\omega \ln(\frac{n_\omega + 1}{n_\omega})$$

$$\implies \Delta S_{\text{photon}} = k_b \ln(\frac{\bar{n}_\omega + 1}{\bar{n}_\omega}). \tag{29.6}$$

At thermal equilibrium the change in entropy must be zero. This gives

$$\Delta S_{\text{photon}} + \Delta S_{\text{semi}} = 0$$

$$\implies k_b \ln(\frac{\bar{n}_\omega + 1}{\bar{n}_\omega}) + \frac{qV - \hbar\omega}{T} = 0 \implies \bar{n}_\omega = \frac{1}{e^{\frac{\hbar\omega - qV}{k_b T}} - 1}, \tag{29.7}$$

and since this spectrum must be identical to Equation 29.4, we obtain

$$a = b = \alpha_{up}(\hbar\omega) \cdot t \cdot e^{\frac{qV - \hbar\omega}{k_b T}}. \tag{29.8}$$

Since the a, b constants remain the same also when the semiconductor is not in thermal equilibrium, the **optical gain** of photons per unit length (units: cm^{-1}) under general conditions is

$$\frac{\Delta n_\omega}{n_\omega t} = \alpha_{up}(\hbar\omega)(e^{\frac{qV - \hbar\omega}{k_b T}} - 1) + \frac{\alpha_{up}(\hbar\omega)}{n_\omega} e^{\frac{qV - \hbar\omega}{k_b T}}$$

$$\implies \boxed{g(\hbar\omega) \approx \alpha_{up}(\hbar\omega)(e^{\frac{qV - \hbar\omega}{k_b T}} - 1).} \tag{29.9}$$

[6] This is the Boltzmann H-theorem discussed in Chapter 21, Section 21.4.

We have neglected spontaneous emission in the gain. The rate of spontaneous emission into mode ω from the semiconductor is

$$\frac{dn_\omega}{dt} = v_g \frac{dn_\omega}{dx} = v_g \frac{a}{t} = v_g \alpha_{up}(\hbar\omega) e^{\frac{qV-\hbar\omega}{k_b T}}, \tag{29.10}$$

and multiplying this by the photon DOS $\frac{n_r^2 \omega^2}{\pi^2 \hbar c^2 v_g}$ gives the spontaneous emission spectrum

$$\boxed{r_{spon}(\hbar\omega) = \frac{n_r^2 \omega^2}{\pi^2 \hbar c^2} \alpha_{up}(\hbar\omega) e^{\frac{qV-\hbar\omega}{k_b T}}.} \tag{29.11}$$

Equations 29.9 and 29.11 emphasize the central importance of the absorption coefficient $\alpha_{up}(\hbar\omega)$ for *both* spontaneous and stimulated emission spectra in semiconductors. Exploiting this, we obtain a relation between the gain spectrum and the spontaneous emission spectrum:

$$\boxed{g(\hbar\omega) = \frac{\pi^2 \hbar c^2}{n_r^2 \omega^2} r_{spon}(\hbar\omega)(1 - e^{\frac{\hbar\omega - qV}{k_b T}}).} \tag{29.12}$$

In Chapter 27 we have discussed absorption coefficients of semiconductors extensively and could proceed and use them to understand spontaneous emission and gain, and the corresponding operation of LEDs and lasers. But we will gain further insight by investigating additional details of how semiconductor bandstructure dictates the nature of the optical transitions. In the next section we show that the two spectra obtained here from considerations of entropy will again be obtained from considerations of the microscopic processes of transitions between individual Bloch states in semiconductor bands.

29.5 Gain and emission in bands

In the entropy method, the gain and spontaneous emission spectra were expressed in terms of the absorption coefficient $\alpha_{up}(\hbar\omega)$ that accounts only for upward transitions in which a photon absorbed by a valence band electron raises it to an empty conduction band state. Here $\alpha_{up}(\hbar\omega)$ depends on the occupation functions f_v and f_c of the valence and conduction band electron states respectively, which in turn depend on the voltage V applied to a semiconductor diode, or on the current I flowing through it. It is useful to express the gain and spontaneous emission spectra in terms of the optical joint density of states $\rho_J(\hbar\omega)$ for light-matter interaction because it is independent of the bias voltage or the injected current. We do so in this section, and in the process identify the connection of the Einstein A and B coefficients of semiconductors with their bandstructure. In spite of our best efforts, we will see that the Einstein A coefficient, which is effectively the inverse of the spontaneous emission lifetime, can only be obtained by its connection to the Einstein B coefficient, and not from direct calculation from the theory of light matter interaction developed till now. Only when the electromagnetic field is fully quantized can the Einstein A-coefficient be calculated directly. We shall postpone the discussion of the process of field quantization to the last sections of this chapter, and discuss lasers and LEDs with the theory of gain and spontaneous emission spectra developed in this section, which is more than sufficient for semiconductor light sources that produce high photon intensities[7].

The **spectra** of spontaneous emission and gain of any semiconductor will now be obtained as a function of the optical joint density of states. Recall from Chapter 27, Equation 27.43 that the $T = 0$ K absorption coefficient spectrum $\alpha_0(\hbar\omega)$ of an **intrinsic semiconductor** is related to its optical joint density of states[8] by

$$\alpha_0(\hbar\omega) = \frac{\pi q^2 |\hat{e}\cdot\mathbf{p}_{cv}|^2}{n_r c \epsilon_0 m_e^2 \omega}\rho_J(\hbar\omega). \tag{29.13}$$

If V is the volume of the semiconductor, $\rho_J(\hbar\omega)Vf_1(1-f_2)$ is the number of states per unit energy available for upward transitions, and $\rho_J(\hbar\omega)Vf_2(1-f_1)$ is the number of states per unit energy available for downward transitions. Considering Fig. 29.8, we denote $f_1 = f_v = 1/(e^{(E_1-F_v)/k_bT}+1)$ as the Fermi–Dirac occupation function of energy eigenvalue E_1 in the valence band with a quasi-Fermi level F_v, and $f_2 = f_c = 1/(e^{(E_2-F_c)/k_bT}+1)$ as the occupation function of an electron eigenvalue in the conduction band with quasi-Fermi level F_c. With these occupation functions, Pauli exclusion principle dictates that for the absorption of a photon of energy $\hbar\omega = E_2 - E_1 = E_{21}$, the initial **k**-state E_1 must be occupied by an electron, and the final state E_2 should be empty. Then the absorption rate (per optical mode volume per unit energy in $1/(\text{cm}^3\cdot\text{eV}\cdot\text{s})$ units) must be $r_{abs} = B_{12}\rho_J(\hbar\omega)Vf_1(1-f_2)L(E_{21})$. Similarly, the stimulated emission rate in the same units is $r_{stim} = B_{21}\rho_J(\hbar\omega)Vf_2(1-f_1)L(E_{21})$, and the spontaneous emission rate is $r_{21}^{spon} = A_{21}\rho_J(\hbar\omega)Vf_2(1-f_2)$. At thermal equilibrium $F_c = F_v$ in the semiconductor, and we must again have detailed balance and the blackbody spectrum of radiation[9]:

$$\begin{gathered} r_{abs} = r_{stim} + r_{spon} \implies \\ B_{12}\rho_J(\hbar\omega)Vf_1(1-f_2)L(E_{21}) = \\ B_{21}\rho_J(\hbar\omega)Vf_2(1-f_1)L(E_{21}) + A_{21}\rho_J(\hbar\omega)Vf_2(1-f_2) \\ \implies L(E_{21}) = \frac{\frac{A_{21}}{B_{21}}}{\frac{B_{12}}{B_{21}}\cdot\frac{f_1(1-f_2)}{f_2(1-f_1)} - 1} = \frac{n_r^3\omega^2}{\pi^2\hbar c^3}\cdot\frac{1}{e^{\frac{\hbar\omega}{k_bT}} - 1} \\ \implies B_{12} = B_{21} = B, \text{ and } \frac{A_{21}}{B_{21}} = \frac{A}{B} = \frac{n_r^3\omega^2}{\pi^2\hbar c^3}. \end{gathered} \tag{29.14}$$

A necessary condition for optical gain is that the **rate of stimulated emission must exceed the rate of absorption**. This is only possible in non-equilibrium. Since the Einstein B coefficient is the same in non-

[7] Of course, field-quantization is necessary to understand *single-photon* sources for quantum communications and qubits for quantum computation, as will be touched upon in the last sections of this chapter after field quantization.

[8] The optical JDOS $\rho_J(\hbar\omega) = \frac{2}{V}\sum_{\mathbf{k}}\delta[E_c(\mathbf{k}) - (E_v(\mathbf{k}) + \hbar\omega)]$ is a function of photon energy that depends only on the semiconductor bandstructure. For example, for 3D parabolic bands of bandgap E_g and no band tail states, it is $\rho_J(\hbar\omega) = \frac{1}{2\pi^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}}\sqrt{\hbar\omega - E_g}$ as was obtained in Chapter 27, Equation 27.44. The corresponding $\rho_J(\hbar\omega)$ for quantum wells, wires, and dots were obtained in Equations 27.46, 27.47, and 27.48 respectively.

[9] In general $\frac{f_1(1-f_2)}{f_2(1-f_1)} = e^{\frac{E_2-E_1-(F_c-F_v)}{k_bT}}$. At thermal equilibrium, $F_c = F_v$ implies $\frac{f_1(1-f_2)}{f_2(1-f_1)} = e^{\frac{E_2-E_1}{k_bT}}$. Also, accounting for the refractive index n_r and speed of light c/n_r of the semiconductor, the blackbody spectrum *inside* the semiconductor is $L(\omega) = \frac{n_r^3\omega^2}{\pi^2\hbar c^3}\cdot\frac{1}{e^{\frac{\hbar\omega}{k_bT}}-1}$, and if n_r changes with photon energy, $n_r^3 \to n_r^3(1 + \frac{\omega}{n_r}\frac{dn_r}{d\omega})$. For simplicity we set $dn_r/d\omega = 0$. Also note the units of the quantities: $r_{ij}^{\prime s} : \frac{1}{\text{cm}^3\cdot\text{s}}$, $A : \frac{1}{\text{cm}^3\cdot\text{s}}$, and $B : \frac{\text{eV}}{\text{s}}$.

equilibrium, we need

$$Bf_2(1-f_1) > Bf_1(1-f_2) \implies e^{\frac{E_1-F_v}{k_bT}} > e^{\frac{E_2-F_c}{k_bT}}$$
$$\implies F_c - F_v \geq E_2 - E_1, \text{ or } \boxed{F_c - F_v \geq \hbar\omega > E_g.} \tag{29.15}$$

The boxed requirement is called the **Bernard–Duraffourg condition** for optical gain in semiconductors after the researchers who formulated them in 1961. From Fig. 29.8, this condition requires the splitting of the quasi-Fermi levels to be larger than the photon energy at which optical gain is desired[10]. When the semiconductor can emit photons of energies larger than the energy gap, optical gain is achieved over the energy window $(E_g, F_c - F_v)$. Clearly, a high density of electrons *and* holes is needed so that $F_c - F_v > E_g$ as well: this is the requirement of **population inversion**. This is achieved either by optical pumping, or by **direct electron and hole current injection in pn junctions**.

The **net** absorption rate per optical mode volume per unit energy in $1/(\text{cm}^3 \cdot \text{eV} \cdot \text{s})$ units is the difference of the upward absorption rate and the downward stimulated emission rate. Thus $r_{abs}^{net} = B\rho_J(\hbar\omega)V[f_1(1-f_2) - f_2(1-f_1)]L(\omega) = B\rho_J(\hbar\omega)V(f_1-f_2)L(\omega)$. This net absorption rate must equal the product of the **net** absorption coefficient $\alpha_0(\hbar\omega)[f_1(1-f_2)-f_2(1-f_1)] = \alpha_0(\hbar\omega)(f_1-f_2)$ and the photon flux $L(\omega)\cdot\frac{c}{n_r}$ which is the density of photons per unit volume per unit energy interval times their group velocity inside the semiconductor. Equating them gives the Einstein B coefficient for the semiconductor, which in turn gives the Einstein A coefficient:

$$B \cdot V\rho_J(\hbar\omega)\cdot(f_1-f_2)L(\omega) = \alpha_0(\hbar\omega)(f_1-f_2)\cdot L(\omega)\cdot\frac{c}{n_r}$$
$$\implies \boxed{B = \frac{\pi q^2|\hat{e}\cdot\mathbf{p}_{cv}|^2}{n_r^2\epsilon_0 m_e^2\omega V} \implies A = \frac{n_r q^2\omega|\hat{e}\cdot\mathbf{p}_{cv}|^2}{\pi\hbar\epsilon_0 c^3 m_e^2 V}.} \tag{29.16}$$

The semiconductor **spontaneous emission spectrum** is therefore

$$r_{spon}(\hbar\omega) = A\rho_J(\hbar\omega)Vf_2(1-f_1) \implies$$
$$\boxed{r_{spon}(\hbar\omega) = \frac{n_r q^2\omega|\hat{e}\cdot\mathbf{p}_{cv}|^2}{\pi\hbar\epsilon_0 c^3 m_e^2}\rho_J(\hbar\omega)f_2(1-f_1) = \frac{n_r^2\omega^2}{\pi^2\hbar c^2}\alpha_0(\hbar\omega)f_2(1-f_1).} \tag{29.17}$$

The net semiconductor **stimulated emission spectrum** is the difference of the downward and upward transitions in the presence of light $r_{stim}(\hbar\omega) = B\rho_J(\hbar\omega)V[f_2(1-f_1) - f_1(1-f_2)]L(\omega) = \frac{c}{n_r}L(\omega)\alpha_0(\hbar\omega)(f_2-f_1)$, giving the **gain spectrum**:

$$\boxed{g(\hbar\omega) = \frac{r_{stim}(\hbar\omega)}{\frac{c}{n_r}L(\omega)} = \frac{\pi q^2|\hat{e}\cdot\mathbf{p}_{cv}|^2}{n_r c\epsilon_0 m_e^2\omega}\rho_J(\hbar\omega)(f_2-f_1) = \alpha_0(\hbar\omega)(f_2-f_1).} \tag{29.18}$$

---

[10] It is *not* necessary for F_c to be inside the conduction band and F_v to be inside the valence band simultaneously. For example if F_v is in the energy gap near the valence band edge E_v, and F_c is well above the conduction band edge E_c, the requirement for gain $F_c - F_v \geq \hbar\omega > E_g$ can be met.

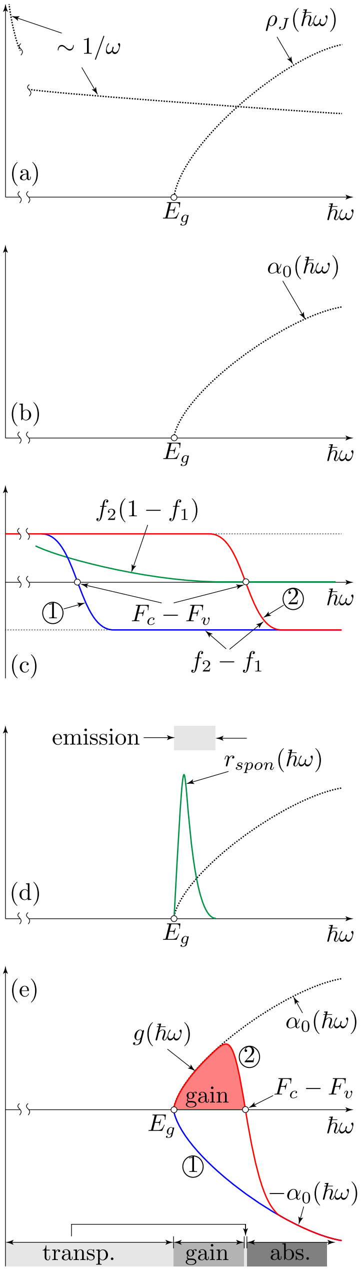


Fig. 29.9 (a) The optical JDOS $\rho_J(\hbar\omega)$, (b) the 0 K intrinsic absorption coefficient $\alpha_0(\hbar\omega)$, and (c) combinations of the conduction and valence occupation functions f_1, f_2 determine (d) the spontaneous emission spectrum $g_{spon}(\hbar\omega)$ (Equation 29.17) and (e) the gain spectrum $g(\hbar\omega)$ (Equation 29.18).

The spontaneous emission spectrum (Equation 29.17) and the gain spectrum (Equation 29.18) and the corresponding forms obtained from the entropy method in Equations 29.11 and 29.12, are identical. This is because in Equations 29.11 and 29.12 the absorption coefficient for only upward transitions is $\alpha_{up}(\hbar\omega) = \alpha_0(\hbar\omega) f_1(1-f_2)$, and $e^{\frac{qV-\hbar\omega}{k_bT}} = \frac{f_2(1-f_1)}{f_1(1-f_2)}$.

In the forms of this section, the two spectra are determined by the semiconductor material parameters via $\rho_J(\hbar\omega)$ (or equivalently $\alpha_0(\hbar\omega)$), and on the state of non-equilibrium (or voltage bias or current injection) via the occupation functions f_1 and f_2. These fundamental optical processes in semiconductors are schematically shown in Fig. 29.9, which is the most important diagram for understanding the operation of LEDs and lasers.

Fig. 29.9 (a) shows the optical joint density of states $\rho_J(\hbar\omega)$ as a function of photon energy $\hbar\omega$. This is a ground-state property of the semiconductor alone and depends on the bandstructure, be it 3D bulk, or quantum wells, wires, or dots. The illustration shows a schematic for 3D bulk, for which $\rho_J(\hbar\omega) \propto \sqrt{\hbar\omega - E_g}$.

Fig. 29.9 (b) shows the 0 K intrinsic absorption coefficient $\alpha_0(\hbar\omega)$, which according to Equation 29.13 goes as $\rho_J(\hbar\omega)/\omega$. It is important to realize that photonic emission processes of interest in the semiconductor occur for a very small window of energies near the energy bandgap E_g, over which $1/\omega$ is nearly constant, as indicated in Fig. 29.9 (a). In this window, the *shape* of $\alpha_0(\hbar\omega)$ is nearly identical to $\rho_J(\hbar\omega)$.

Fig. 29.9 (c) shows the factors due to occupation functions for the spontaneous emission spectrum and two cases labeled ① and ② for the gain spectrum. From Equation 29.17, the **factor** $f_2(1-f_1)$ **is always positive for spontaneous emission**. It has a tail that extends above $\hbar\omega = E_g$. This tail multiplies with $\propto \omega^2\alpha_0(\hbar\omega)$ to produce a narrow window of photon energies near the bandgap shown in Fig. 29.9 (d), which is the spontaneous emission spectrum. This spectrum is the output of a semiconductor LED.

The optical gain spectrum according to Equation 29.18 depends on the product of the factor $f_2 - f_1$ with $\alpha_0(\hbar\omega)$. In Fig. 29.9 (c) the line labeled ① is $f_2 - f_1$ when the Bernard–Duraffourg condition is not met, and $F_c - F_v < E_g$. In that case $f_2 - f_1 \approx -1$ for $\hbar\omega > E_g$, and the gain spectrum shown in Fig. 29.9 (e) is simply $-\alpha_0(\hbar\omega)$. For the case labeled ②, the Bernard–Duraffourg condition is indeed met because $F_c - F_v > E_g$. **In that case** $f_2 - f_1 > 0$ **in a small photon energy window** $E_g \leq \hbar\omega \leq F_c - F_v$ **above the bandgap** E_g**, and the gain becomes positive**. Within this window of energies, stimulated emission produces optical gain, and powers the operation of semiconductor optical amplifiers and lasers.

Several additional features of importance for LEDs and lasers can be pointed out in Fig. 29.9. Band-tail states due to inhomogeneities broaden the sharp band edges of $\rho_J(\hbar\omega)$ and $\alpha_0(\hbar\omega)$, which in turn broaden the $r_{spon}(\hbar\omega)$ and $g(\hbar\omega)$, but leave the prior analysis intact.

For a semiconductor with population inversion such that $F_c - F_v > E_g$, a photon of energy $\hbar\omega = F_c - F_v$ experiences a net zero gain or loss (Fig. 29.9 (e)). In other words the semiconductor is transparent to this photon, even though the photon energy is larger than the bandgap! This is called **induced transparency**. For photon energies $\hbar\omega > F_c - F_v$, the gain becomes negative and they are absorbed. It is important to note that under bias conditions such that $F_c - F_v > E_g$, spontaneous emission still occurs in the background, though the number of photons created by stimulated emission far outnumber them. We are now ready to discuss LEDs and lasers in the following sections.

29.6 Spontaneous emission: LEDs

Semiconductor LEDs are primarily pn junction diodes as indicated in Fig. 29.10 (a). The quasi-Fermi levels F_c and F_v change according to the Shockley relation $F_c - F_v = qV$ as was discussed in Chapter 18. The corresponding excess electron and hole densities injected into the depletion region (also discussed in Chapter 18) are responsible for the movement of the F_c closer to the conduction band, and F_v closer to the valence band (Fig. 29.10 (b)). In most modern LEDs, quantum wells or quantum dots are inserted in the depletion region to tune the wavelength of emission, and also to increase the oscillator strength of the optical transition. In Equation 29.19 one must then use the corresponding $\alpha_0(\hbar\omega)$ which depend on the joint DOS $\rho_J(\hbar\omega)$ of such quantum confined structures. Several LED design parameters are independent of the details of the active region design, so we discuss them assuming a bulk semiconductor next.

From Equation 29.17 we get the spontaneous emission spectrum

$$r_{spon}(\hbar\omega) = \frac{n_r^2\omega^2}{\pi^2\hbar c^2} \frac{e^{\frac{E_1-F_v}{k_bT}}}{(e^{\frac{E_2-F_c}{k_bT}}+1)(e^{\frac{E_1-F_v}{k_bT}}+1)} \alpha_0(\hbar\omega)$$

$$\text{low level injection} \implies \boxed{r_{spon}(\hbar\omega) \approx \frac{n_r^2\omega^2}{\pi^2\hbar c^2} e^{-\frac{\hbar\omega}{k_bT}} e^{\frac{qV}{k_bT}} \alpha_0(\hbar\omega).} \quad (29.19)$$

Fig. 29.10 (c) shows how the spontaneous emission rate depends on the absorption coefficient and the applied voltage. For the approximate form in Equation 29.19, we use $E_2 - E_1 = \hbar\omega$, the criteria $\hbar\omega >> k_bT$ which is typical for photons in the IR, visible, and UV wavelengths at room temperature, and low-level carrier injection conditions $\frac{E_2-F_c}{k_bT} >> +1$ and $\frac{E_1-F_v}{k_bT} << -1$, and that the quasi-Fermi level splitting $F_c - F_v = qV$ is small compared to the bandgap of the pn junctions. The fact that the spontaneous emission rate is proportional to the semiconductor absorption coefficient is called the **Roosbroeck–Shockley relation** for semiconductors.

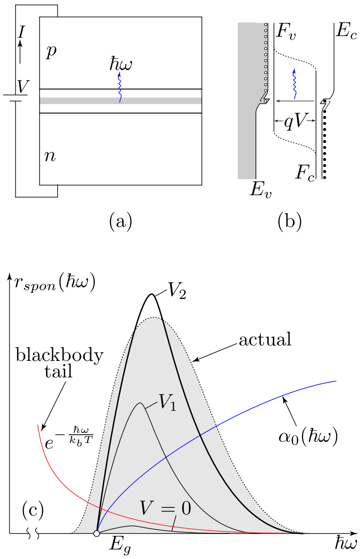


Fig. 29.10 Schematic diagram showing the spontaneous emission spectrum as a function of the splitting of the quasi-Fermi level $F_c - F_v = qV$ for $V_2 > V_1$, and the relation of the spontaneous emission spectrum to the blackbody spectrum and the ideal equilibrium absorption coefficient $\alpha_0(\hbar\omega)$ of the semiconductor. The actual spectrum (say at $F_c - F_v = qV_2$) is broadened due to band tail states and inhomogeneities.

Fig. 29.11 Oleg Losev was one of the earliest inventors of the semiconductor LED.

At thermal equilibrium when $V = 0$ and $F_c = F_v$, Equation 29.19 states that the emission spectrum of the LED goes as $\propto \omega^2 e^{-\frac{\hbar\omega}{k_bT}} \alpha_0(\hbar\omega)$. This is a product of the photon DOS, the (tail of the) blackbody spectrum, multiplied by the absorption coefficient of the semiconductor, as indicated schematically in Fig. 29.10 (c). The spontaneous emission rate increases exponentially with voltage as e^{qV/k_bT}. Another way to state this is to write the radiative recombination rate as $1/\tau_r = \mathcal{B}np$, where $\mathcal{B}$ is a radiative recombination constant, and the product of the excess injected electron concentrations n and hole concentration p is due to the occupation factor $f_1(1 - f_2)$. Carriers injected into the LED active region may be lost[11] to non-radiative Shockley–Read–Hall type processes (Fig. 29.12), for which $1/\tau_{nr,srh} = \mathcal{A}n$, or to the non-radiative Auger recombination at a rate $1/\tau_{nr,Auger} = \mathcal{C}n^3$. Then, the IQE or **internal quantum efficiency** η_{iqe} of a semiconductor LED is defined as the ratio of photons generated per unit time $\mathcal{N}$ to the rate of electron injection into the active region I/q. The IQE then is

$$\eta_{iqe} = \frac{\mathcal{N}}{I/q} = \frac{1/\tau_r}{1/\tau_r + 1/\tau_{nr,srh} + 1/\tau_{nr,Auger}} = \frac{\mathcal{B}n^2}{\mathcal{A}n + \mathcal{B}n^2 + \mathcal{C}n^3}, \tag{29.20}$$

which is called the $\mathcal{ABC}$ model of the LED IQE. Clearly the goal is to increase $\mathcal{B}$ and reduce $\mathcal{A}$ and $\mathcal{C}$ as far as feasible. Since the pn junction LED emits light in the forward bias, the net output power in watts is

$$P_{out} = \eta_{extraction} \cdot \mathcal{N} \cdot \hbar\omega = \underbrace{\eta_{extraction} \cdot \eta_{iqe}}_{\eta_{tot}} \cdot \frac{\hbar\omega}{q} \cdot I = \eta_{tot} \frac{1.24}{\lambda} I. \tag{29.21}$$

Of the $\mathcal{N}$ photons per second emitted in the active region, a fraction $\eta_{extraction}$ are extracted from the semiconductor, the rest are reflected back due to the difference in the refractive index. Hence, the total power output is $P_{out} = \eta_{tot} \frac{1.24}{\lambda} I$ where the wavelength λ is expressed in microns, or $P_{out}/I = \eta_{tot} \frac{1.24}{\lambda}$ watts/ampere. Note that this is the inverse of the photodetector metric of responsivity R_λ in A/W units (Chapter 28). The output power is linearly proportional to the injected current as long as η_{tot} does not depend on I. Typically for LEDs $I/A = 1 \leq J \leq 50$ A/cm^2. For example, in a blue LED with $\lambda = 450$ nm, $\eta_{tot} = 0.6$, and injection current $I = 10$ mA produces $P_{out} = 16.5$ mW output power. Generally at high injection current densities, η_{tot} reduces due to Auger effects, and there is "droop" in the LED efficiency. What this efficiency hides is the net efficiency of power conversion from electrical to optical. The wall-plug efficiency quantifies that:

$$\eta_{WPE} = \frac{P_{out}}{P_{in}} = \frac{\eta_{extraction}\eta_{iqe}\frac{\hbar\omega}{q} \cdot I}{I \cdot V} = \eta_{extraction}\eta_{iqe}\frac{\hbar\omega}{qV}, \tag{29.22}$$

which, if for the same blue LED ($\lambda = 450$ nm, or $\hbar\omega = 2.76$ eV) a voltage $V = 3.5$ Volt is needed to drive 10 mA of current, leads

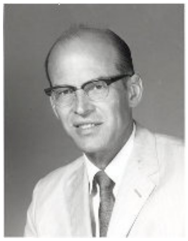


Fig. 29.12 Robert N Hall (the H of the SRH) was a scientist at General Electric. He contributed to the development and understanding of pn junctions and tunnel diodes, and was instrumental in the demonstration of the first semiconductor pn junction LEDs and lasers.

[11] The general definition of the absorption coefficient can account for band-tails and inhomogeneous broadening by defects by introducing a finite linewidth Γ of transitions in the Lorentzian form $\delta[E_b - (E_a + \hbar\omega)] \rightarrow \frac{\Gamma/2\pi}{[E_b - (E_a + \hbar\omega)]^2 + (\Gamma/2)^2}$ such that $\int \delta[E_b - (E_a + \hbar\omega)] d(\hbar\omega) = 1$. This can account for the actual broadened LED output spectrum of Fig. 29.10 (b) compared to that for the ideal semiconductor active region without band tails.

to $\eta_{WPE} = 0.6 \times \frac{2.76}{3.50} = 0.41$, or a 41% wall-plug efficiency, a typical number for low-injection blue LEDs used in solid-state lighting today.

Semiconductor LEDs find a very wide range of uses today. The highest volume usage of LEDs are in displays (monitors, screens, billboards) and in lighting (indoor, automotive). For such needs, red LEDs based on InGaAsP, and blue and green LEDs based on InGaN semiconductor pn diodes cover the visible spectrum. UV and deep-UV LEDs based on AlGaN semiconductors are increasingly being used for polymer curing, disinfection, and in advanced diagnostic and therapeutic applications. In the IR wavelengths InGaAs based LEDs are used in low data rate communication systems. Several methods are used to increase the efficiency of semiconductor LEDs. The primary techniques used for light extraction are roughening of the surface to prevent total internal reflection and by the use of anti-reflection coatings. A typical LED output spectrum has a full-width at half maximum (FWHM) of $\sim 10-20$ nm in wavelength around the peak emission wavelength. The photons are incoherent and not directional, because they are generated by the spontaneous emission process.

The spontaneous emission lifetime restricts the speed of modulation of LEDs to below 1 GHz. Therefore they are used for low-data rate communications (see Fig. 29.13). Since the LED output is not directional, special lenses must be used to focus and couple the light output from the LED into optical waveguides.

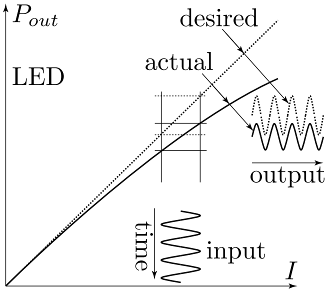


Fig. 29.13 Illustration of the use of a LED for communication systems, and the limitation of modulation and linearity. Because of the decrease in efficiency at high injection currents, when a time-varying input current is applied, the output power modulation is reduced. In addition the non-linear output power causes interference of input signals of different frequencies. The modulation speed of the LED is restricted by the rate of spontaneous emission. These limitations of the LED as a source for communication systems are solved by the laser.

At high injection currents into the LED, the Bernard–Duraffourg condition for optical gain can be reached, triggering amplification of the spontaneously emitted photons. LEDs in which such **amplified spontaneous emission** is used are called **superluminescent LEDs** or SLEDs. Their properties are in-between those of a semiconductor LED and a semiconductor laser diode. For example, the emission spectrum becomes narrower and the output power P_{out} vs. I becomes superlinear because of the gain process. However, because of the absence of a feedback mechanism (= mirrors), they do not go into laser oscillations. That is the topic of the next section.

29.7 Stimulated emission: lasers

As the injection current in a semiconductor pn junction is increased, the minority electron and hole densities injected into the optically active region increase exponentially with voltage. The spontaneous emission rate increases, but so does the stimulated emission rate (see Fig. 29.14). From Equations 29.17 and 29.18, the ratio of the optical gain to the spontaneous emission is

$$\boxed{\frac{g(\hbar\omega)}{r_{spon}(\hbar\omega)} = \frac{\pi^2\hbar c^2}{n_r^2\omega^2}(1 - e^{\frac{\hbar\omega - qV}{k_b T}})} > 0 \text{ for } qV = F_c - F_v > \hbar\omega. \quad (29.23)$$

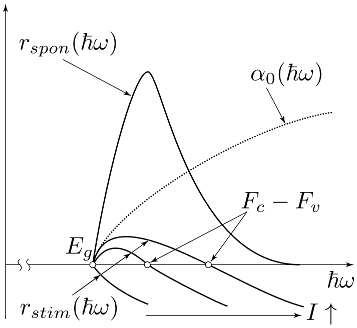


Fig. 29.14 The stimulated emission rate changes from negative (= photon absorption) to positive (= photon emission) with increasing current injection. The relative magnitudes of the stimulated rates should be compared among themselves, but not quantitatively with the spontaneous rate in this illustration which is shown for the purpose of qualitative comparison. The spontaneous emission rate which is always positive is shown here for a very high bias condition.

When the quasi-Fermi level separation $F_c - F_v = qV$ exceeds $\hbar\omega$, the Bernard–Duraffourg inversion condition is met. Then for photons of

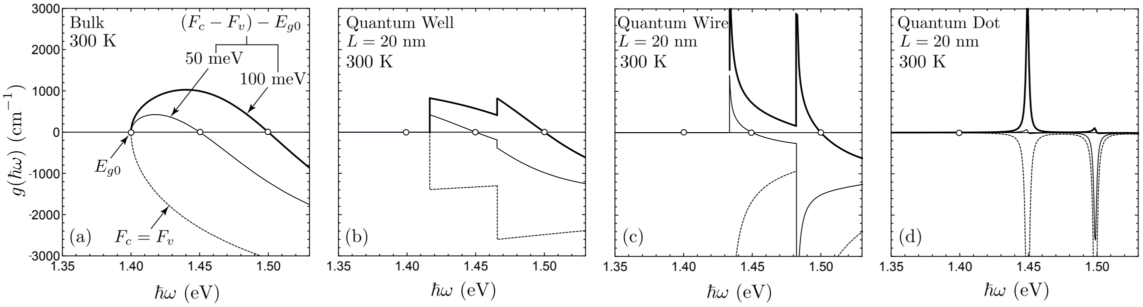


Fig. 29.15 Calculated 300K gain spectra of semiconductors in (a) 3D bulk, (b) 2D quantum wells, (c) 1D quantum wires, and (d) 0D quantum dots for equilibrium ($F_c = F_v$, dashed line), above population inversion at $F_c - F_v = E_{g0} + 50$ meV (thin solid line) and above population inversion at $F_c - F_v = E_{g0} + 100$ meV (thick solid line), with a 80%-20% split between $F_c - E_{c0}$ and $E_{v0} - F_v$. The dashed lines are $g(\hbar\omega) = -\alpha_0(\hbar\omega) \propto -\rho_J(\hbar\omega)/\omega$. The semiconductor bulk band parameters are $E_{g0} = 1.4$ eV, $m_c^\star = 0.067 m_e$, $m_v^\star = 0.4 m_e$, refractive index $n_r = 3.3$, and the Einstein A coefficient of $A = 1$ ns^{-1}. Though representative of GaAs, band tails and multiple valence bands are neglected except for the quantum dot for which a broadening of $\Gamma \sim E_{g0}/600$ is assumed. All quantized axes are of length $L = 20$ nm, and an infinite barrier approximation is used. Exciton effects are neglected.

energy $E_g \leq \hbar\omega \leq F_c - F_v$, $\exp[(\hbar\omega - qV)/(k_b T)] < 1$ and the gain becomes positive. Another way to look at the same thing is to say that the **net absorption coefficient** $\alpha(\hbar\omega) = \alpha_0(\hbar\omega)(f_1 - f_2)$ **becomes negative**. From the Beer–Lambert law, in the spectral window within which the absorption coefficient becomes negative, the intensity of light $I(z) = I_0 e^{-\alpha(\hbar\omega)z}$ actually **increases** as the light propagates within the semiconductor medium, as was indicated in Fig. 16.20. This remarkable property is **optical gain** because more photons exit than enter the semiconductor. The semiconductor medium that exhibits this non-equilibrium condition is therefore a **gain medium**, one of the requirements of a laser. The energy for the additional photons comes from the excitation source that maintains the population inversion – for example, the steady state electrical injection of excess carriers.

Fig. 29.15 shows how the optical gain spectrum for 3D bulk semiconductors and for quantum wells, wires, and dots changes with increasing current injection. The gain coefficient starts out at $F_c - F_v = 0$ as a "loss-coefficient", or the absorption coefficient $\alpha_0(\hbar\omega)$ which closely follows the characteristic optical joint density of states $\rho_J(\hbar\omega)$ of the bulk, and quantized wells wires and dot heterostructures, as indicated by the dashed lines. Quantum confinement also shifts the band-edges compared to E_{g0}, that of the bulk semiconductor. Thus the bulk gain spectrum has a single maximum whereas that of the quantized heterostructures may have multiple maxima at high $F_c - F_v$. Note the values of the gain coefficients evaluated here for $F_c - F_v = E_{g0} + 100$ meV is of the order 1000 cm^{-1}, a typical number. We will see shortly that only a fraction of light characterized by an overlap parameter Γ actually experience this gain. This is why the gain of a semiconductor laser is typically reported as $\Gamma g(\hbar\omega)$ rather than $g(\hbar\omega)$ alone. Fig. 29.16 shows the dependence of gain on temperature for a 3D bulk semiconductor for a lower injection than Fig. 29.15 (a).

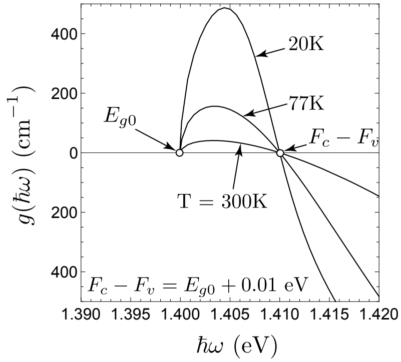


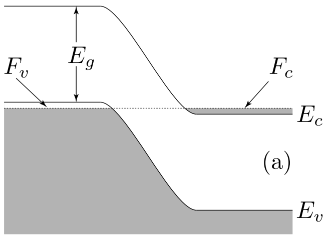
Fig. 29.16 Calculated temperature-dependence of the optical gain. $F_c - F_v = E_g + 0.01$ eV with $F_c = E_g + 0.007$ eV and $F_v = -0.003$ eV with the valence band edge assumed to be the zero of energy. The same material parameters as Fig. 29.15 are used. The maximum gain increases at lower temperatures, but the transparency photon energy $\hbar\omega = F_c - F_v$ does not change. The dependence of the gain coefficient on temperature leads to undesirable thermal changes in the laser output characteristics (see Exercise 29.7).

29.8 Double heterostructure lasers

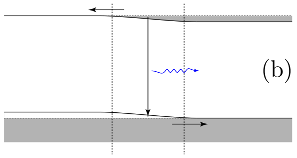
Figures 29.17 (a – f) show the evolution of the design of active regions of semiconductor laser diodes over the years. The earliest laser diodes were degenerately doped pn homojunctions of the direct bandgap semiconductors based on GaAs or InGaAs. The energy band diagram of such a diode is shown in Fig. 29.17 (a) when $V = 0$, and in Fig. 29.17 (b) when a high forward bias voltage is applied. Because of degenerate doping, the quasi-Fermi levels $F_c > E_c$ on the n-side, and $F_v < E_v$ on the p-side, and $F_c = F_v$ everywhere at equilibrium. But at a large forward bias, significant amount of electrons and holes are injected into the depletion region. The population inversion condition $qV = F_c - F_v > \hbar\omega > E_g$ is achieved in a small window inside the depletion region as seen in in Fig. 29.17 (b). This small region has optical gain. When an optical cavity is formed around this region (e.g. as in Fig. 29.18), and the injection is enough to overcome the optical losses in this cavity, lasing is achieved by direct electrical injection.

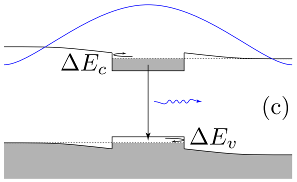
The earliest lasers of this kind achieved in the early 1960s could only be operated at low temperatures, and under **pulsed** conditions when the injection current was applied in pulses to avoid overheating. Room temperature operation, as well as the laser oscillations under **continuous wave** (CW) injection was not possible in such structures. One of the reasons is seen in Fig. 29.17 (b): a large portion of electrons and holes injected simply overshoot the depletion region to enter the other side of the junction as indicated by arrows. These carriers do not contribute to optical gain, but do contribute to Joule heating via $I \cdot V$. Furthermore, the optical gain is over a small region inside the depletion region. The optical mode is rather large, of the order of λ_0/n_r where $\lambda_0 \approx 2\pi\hbar c/E_g$ is the free space wavelength , in 100s of nm to 1 μm. Only a small portion of it, about ~ 50 nm experiences optical gain, while a significant portion of it not just does not experience optical gain, it experiences significant loss due to free-carrier absorption owing to its overlap with the degenerately doped n and p-regions.

To solve this problem, Herbert Kroemer (Fig. 15.21) in 1963 proposed the use of a **double heterostructure** as shown in Fig. 29.17 (c). The band offsets ΔE_c and ΔE_v now prevent carriers from overshooting, and rather dramatically lower the required threshold injection current for lasing. This approach led to the successful demonstration of room-temperature and continuous wave lasing around 1970 by multiple groups. The double-heterostructure also offers another significant benefit: **a semiconductor of a smaller bandgap also has a larger refractive index**. The refractive index difference results in a **natural optical waveguide**, keeping the optical mode close to the gain medium. The reduced overlap of the optical mode with the p- and n-regions lowers the free-carrier absorption loss somewhat. The **separate confinement heterostructure** (SCH) in Fig. 29.17 (d) lowers the free-carrier absorption losses even further by introducing **clad** layers without doping and an intermediate bandgap to provide separate cav-

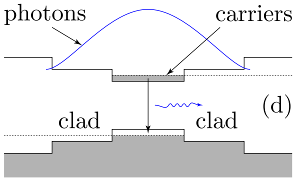


pn junction

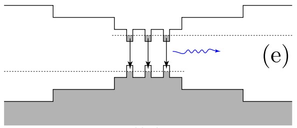




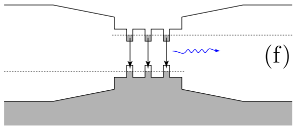
double heterostructure



separate confinement heterostructure



quantum well laser



GRINSCH

Fig. 29.17 Evolution of the design of semiconductor lasers.

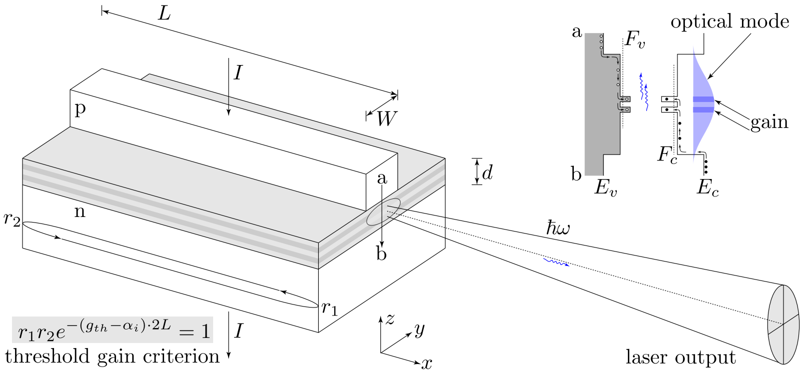


Fig. 29.18 Construction and output of an edge emitting semiconductor diode laser.

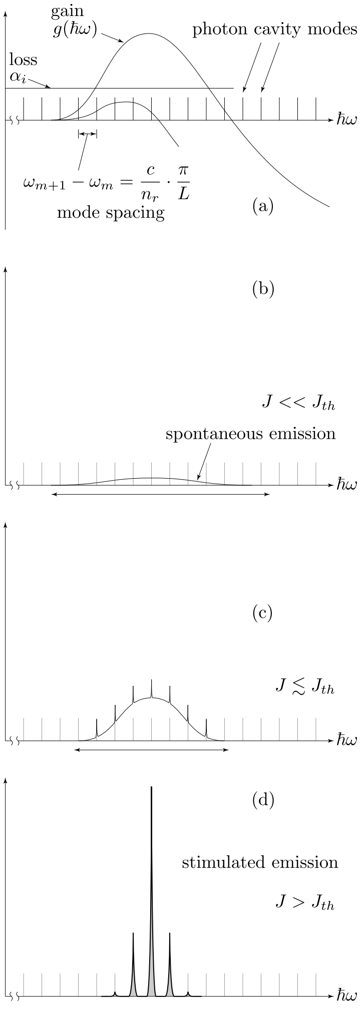


Fig. 29.19 Laser modes and operation. Not exact, meant for illustration. The stimulated emission spectrum axis in (d) spans much larger values than (b, c).

ities for photons and carriers. The active region can consist quantum wells (Fig. 29.17 (e)) or quantum dots, and the clad layers can be of a graded energy bandgap. The resulting graded refractive index gives them the name GRINSCH lasers.

Fig. 29.18 shows the physical construction of a **edge-emitting** SCH quantum well laser. It uses a Fabry–Perot cavity to select modes such that the laser output emerges from the edge, as shown. The semiconductor heterostructure is typically grown epitaxially by MBE or MOCVD. In the geometry shown, starting from the substrate are the wider bandgap n-type layer, the clad+carrier confinement region with quantum wells, followed by the wider-bandgap p-type layer on the surface. The energy band diagram, carrier injection, and optical mode with the regions of optical gain under population inversion are indicated. Note that only a fraction of the optical mode indicated by the stripes experence optical gain. The planar structure is processed after growth into mesas of typical widths $W \sim 5 - 10$ μm and cavity length $W \sim 200 - 500$ μm by etching for the sidewalls around the width, and cleaving the ends along the length for a smooth, mirror-finish. The cleaved facets create optical reflectivities of r_1 and r_2, forming a planar cavity of length L. In some embodiments the mesa may be etched down to the n-layer. Typically exposed surfaces near the carrier flow path are passivated to prevent the loss of carriers to non-radiative surface recombination processes. The current is injected from low-resistance ohmic contacts from the top and collected at the bottom as indicated by arrows. When the current exceeds a threshold $I > I_{th}$, a laser beam emerges from the side as indicated.

Lasing conditions: We now evaluate the fundamental requirements to achieve laser operation. Let all non-radiative recombination processes in the optical waveguide lead to a loss of propagating photons characterized by an absorption coefficient α_i as shown in Fig. 29.19 (a). The gain spectrum $g(\hbar\omega)$ corresponding to a specific injected current depends on the splitting $qV = F_c - F_v$ achieved at that current. For currents less than the threshold for lasing $I < I_{th}$, the gain spectrum is positive and population inversion is achieved, but the maximum gain falls below the loss α_i. Therefore the threshold current I_{th} is the current for which the maximum gain g_{max} exceeds loss α_i. The current density at which the gain g becomes positive but is still below α_i is called the **transparency current** I_{tr}, where clearly $I_{th} > I_{tr}$.

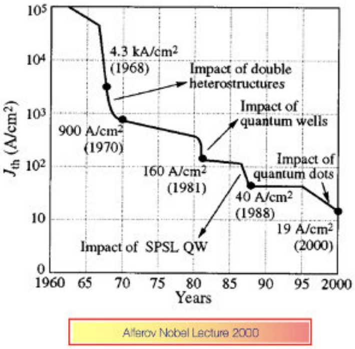


Fig. 29.20 Lowering of threshold current of semiconductor lasers over the years by introduction of double heterostructures, quantum wells, and quantum dots. Adapted from Reviews of Modern Physics 73, 767 (2001).

Because the electromagnetic field must go to zero at the mirrors located along the x-axis , the boundary condition dictates $\sin(k_x L) = 0 \implies k_x L = m\pi$ where m is an integer. Thus the allowed electromagnetic modes $k_m = m\frac{\pi}{L}$ are separated by $k_{m+1} - k_m = \frac{\pi}{L}$, which with $\omega_m = \frac{c}{n_r} k_m$ leads to a mode spacing $\omega_{m+1} - \omega_m = \frac{c}{n_r}\frac{\pi}{L}$ as indicated in Fig. 29.19 (a). For a cavity length $L >> \lambda_0$ where λ_0 is the free space wavelength of the laser output, the mode spacing $\frac{\pi}{L}$ is small enough that a large number of allowed modes fit in the small gain window within which $g > \alpha_i$. The laser can produce stimulated emission output in any or all of these modes.

The current density $J = I/A$ is nominally independent of the size. Figures 29.19 (b–d) show how the emission evolves from spontaneous emission for $J << J_{th}$ to the combination of spontaneous and stimulated emission for $J \lesssim J_{th}$ to the dominant stimulated emission for $J > J_{th}$. This transition is accompanied by a significant narrowing of the spectrum in $\hbar\omega$ or in wavelengths, and in a very large increase in the emission intensity of the primary laser modes, since photon emission changes from into $\sim 10^{10}$ modes in spontaneous emission to $\sim 1 - 100$ in stimulated emission, depending on the type of the laser.

Consider the active region of the laser is at gain g due to current injection. Then the intensity ($\propto |E|^2$) of an electromagnetic wave propagating in the x-direction along the length L of the optical cavity varies as $I(x) = I(0)e^{gx}e^{-\alpha_i x}$. Since the reflectivities of the two mirrors are r_1 and r_2, the net change in amplitude for a round-trip of length $2L$ is $r_1 r_2 e^{(g-\alpha_i)\cdot 2L}$. The requirement for the laser threshold is that the field replicates itself after a round trip such that **gain = loss**. This condition translates to the required threshold gain[12]

$$r_1 r_2 e^{(g_{th}-\alpha_i)\cdot 2L} = 1 \implies \boxed{g_{th} = \alpha_i + \frac{1}{2L}\ln(\frac{1}{r_1 r_2}).} \quad (29.24)$$

As a typical example for semiconductor lasers, assuming the refractive index of the semiconductor to be $n_r = 3.5$, $r_1 = r_2 = \frac{n_r - 1}{n_r + 1} = 0.56$ and $R = r_1 r_2 = 0.31$, fairly typical for IR lasers. Assuming a typical loss coefficient of $\alpha_i = 10$ cm^{-1} for a cavity length $L = 250$ μm, the required threshold gain is $g_{th} = 10 + \frac{1}{0.05}\ln(\frac{1}{0.31}) \approx 33.5$ cm^{-1}. It is important to note that only the fraction of photons that overlap with the gain medium experience gain. Using this overlap factor $\Gamma = \frac{n_{gain}}{n_{eff}} \cdot \frac{\int_{gain} dz \cdot |E(z)|^2}{\int_{-\infty}^{+\infty} dz \cdot |E(z)|^2}$ this actual required threshold gain is given by $g_{th}^{req} = g_{th}/\Gamma$. For a typical $\Gamma \sim 0.04$ for edge-emitting double heterostructures, the actual required gain coefficient in the region of population inversion becomes $g_{th}^{req} \approx 837.8$ cm^{-1}. This is a typical required gain for semiconductor lasers[13]. Typical injected current densities of $J \approx 200$ A/cm^2 can provide this gain at room temperature (Fig. 29.20).

[12] The total mirror loss $\frac{1}{2L}\ln(\frac{1}{r_1 r_2})$ is denoted by α_m, yielding the threshold material gain $g_{th} = \alpha_i + \alpha_m$ for the following sections. Note typical orders of magnitude of $\alpha_i \sim 10$ cm^{-1} and $\alpha_m \sim 20$ cm^{-1} for edge-emitting lasers with cleaved facet mirrors.

[13] Note from Fig. 29.15 that peak optical gain of this magnitude is available, for example, for $F_c - F_v \sim E_{g0} + 100$ meV. Also, the factor n_{gain}/n_{eff} in Γ is the ration of refractive indices between the gain region and the effective refractive index of the waveguide.

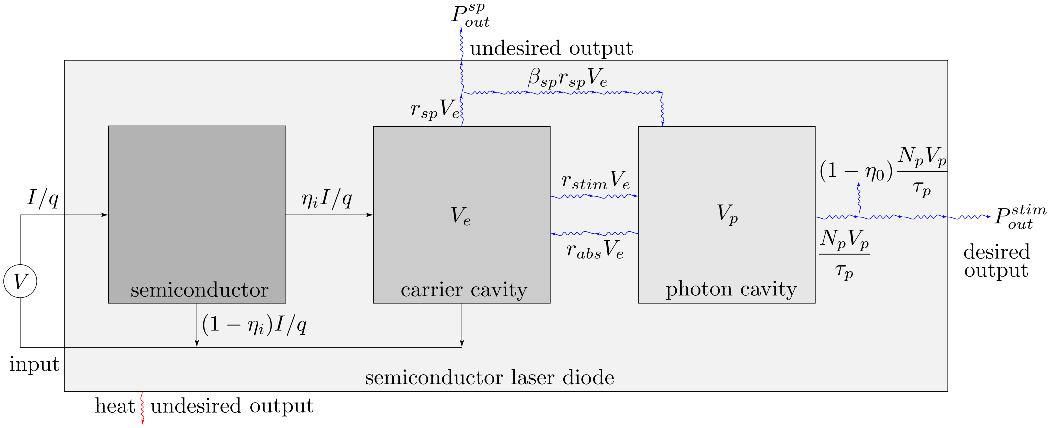


Fig. 29.21 A block diagram for a phenomenological model of a semiconductor laser diode that connects the desired (and undesired) outputs to the input. Energy provided on the left as input via a current I and voltage V across the pn diode produces the desired stimulated emission power P_{out}^{stim}. The undesired outputs are spontaneous emission of power P_{out}^{spon} and heat.

29.9 Laser rate equations

The microscopic and quantum mechanical details of the semiconductor laser of the prior sections guide the design and understanding of the semiconductor laser. But its operation and use is captured in a rather simple phenomenological model of rate equations which rests on the shoulders of the microscopic details. We briefly discuss the rate equation model in this section and understand it's predicted output characteristics. Then in the following section, a case-study of a GaN based blue-laser diode is presented, which exhibits characteristics of both the microscopic model and that of the rate equation model.

Fig. 29.21 shows a block diagram of a semiconductor laser diode. The input current I flowing into the device pumps carriers at the rate I/q. These carriers set a number of microscopic processes in motion inside the device. The end result of all microscopic processes are three macroscopic outputs shown as wavy arrows exiting the rectangular box: 1) desired stimulated emission of power P_{out}^{stim}, 2) undesired background spontaneous emission P_{out}^{sp}, and 3) undesired heat dissipated[14]. The following phenomenological equation tracks injected carriers:

$$\frac{d(NV_e)}{dt} = \underbrace{\eta_i \frac{I}{q}}_{\text{gen.}} - \underbrace{\frac{NV_e}{\tau}}_{\text{spon. + non-rad. rec.}} - \underbrace{v_g \cdot g \cdot N_p V_e}_{\text{stim. rec.}}. \tag{29.25}$$

A fraction $\eta_i I/q$ of carriers injected from the diode terminals fall into the "carrier cavity" of volume[15] V_e. The remaining component $(1-\eta_i)I/q$ does not produce desired optical output. The NV_e electrons and holes trapped in the carrier cavity recombine by spontaneous emission and non-radiative recombination at the rate $1/\tau$. The stimulated emission recombination rate $v_g \cdot g \times (N_p V_e)$ is proportional to the photons

[14] The heat generated is due to Joule heating, emission of phonons by electron/hole relaxation, and non-radiative recombination, which are all interrelated! In the following discussion, N is the volume density of carriers and NV_e is the *number* of carriers in volume V_e. Similarly N_p the volume density of photons and $N_p V_p$ is the *number* of photons in volume V_p, and $N_p V_e$ is the number of photons in volume V_e.

[15] This carrier cavity is the narrower bandgap region in a double heterostructure laser which may consist of quantum wells or dots. Importantly, this is the volume in which optical gain $g(\hbar\omega)$ is available when the Bernard–Duraffourg criteria $F_c - F_v > \hbar\omega > E_g$ is met.

present in the *carrier cavity*, where v_g is the light group velocity and g is the *net* optical gain which includes stimulated emission + absorption.

The fate of the desired cavity photons is tracked by

$$\frac{d(N_pV_p)}{dt} = \underbrace{+v_g \cdot g \cdot N_pV_e}_{\text{gen., stim.}} + \underbrace{\beta_{sp}R_{sp}V_e}_{\text{gen., spon.}} - \underbrace{\frac{N_pV_p}{\tau_p}}_{\text{lost, output}} \quad . \qquad (29.26)$$

Photons emitted by stimulated emission go into desired mode(s) allowed in the photon cavity of volume V_p, typically of thickness $\sim \lambda_0/2n_r$ for edge-emitting lasers, which is larger than V_e. But since optical gain is available only in the carrier cavity, only a fraction $\Gamma = V_e/V_p$ of the photons are amplified. A small fraction $\beta_{sp} \sim 10^{-6}$ of photons produced by spontaneous emission at the rate R_{sp} in volume V_e go into the desired cavity modes of the laser and the rest escape as undesired output P_{out}^{sp}. Finally, the desired photons are lost from the cavity at the rate $\frac{1}{\tau_p} = v_g(\alpha_i + \alpha_m)$, which is why τ_p is the **photon lifetime** in the cavity. The fact that τ_p is only a property of the mirrors and losses and *not* of the injected carrier current is seen clearly by putting the generation terms of Equation 29.26 to zero, in which case $\frac{dN_p}{dt} = -\frac{N_p}{\tau_p}$ yields $N_p(t) = N_p(0)e^{-t/\tau_p}$, also justifying the name of a lifetime. Equations 29.26 and 29.25 then become (Equations 29.27 (a) and (b)):

$$\boxed{\frac{dN}{dt} = \eta_i\frac{I}{qV_e} - \frac{N}{\tau} - v_g g N_p} \text{ and } \boxed{\frac{dN_p}{dt} = \Gamma v_g g N_p + \Gamma\beta_{sp}R_{sp} - \frac{N_p}{\tau_p},} \qquad (29.27)$$

which are the **laser rate equations**. With the microscopic insights discovered earlier, these all-important rate equations can now be used to explain how the processes of Fig. 29.21 lead to the steady state *and* dynamical output characteristics of lasers shown in Fig. 29.22.

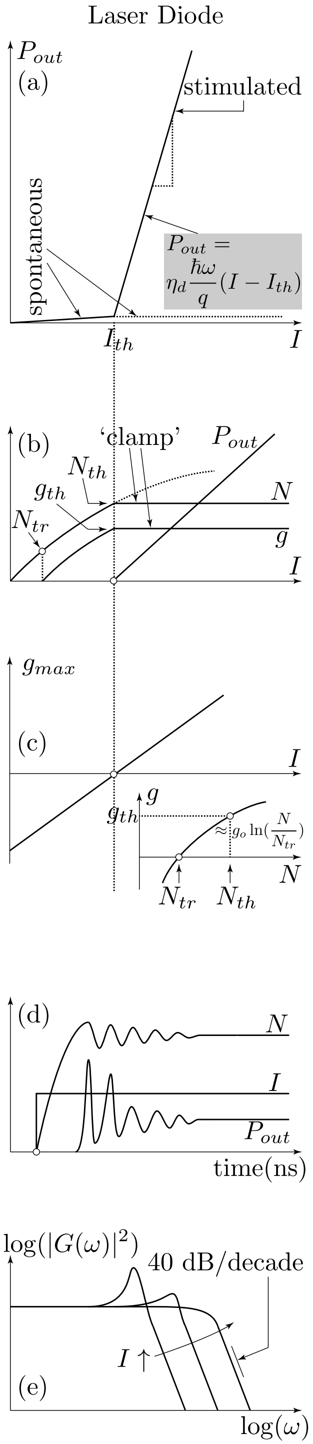


Fig. 29.22 Phenomenological model of semiconductor diode lasers.

Steady-state behavior: Consider the steady state ($d(...)/dt = 0$) laser output power vs. input current characteristics of Fig. 29.22 (a). At $I \approx I_{th}$ stimulated emission may be neglected, giving $N_{th}/\tau = \eta_i I_{th}/(qV_e)$. For $I > I_{th}$ above threshold, the carrier density $N = N_{th}$ and gain $g = g_{th}$ **clamp at the threshold value** as shown in Fig. 29.22 (b). Here $\Gamma g_{th} = \alpha_i + \alpha_m$ defines the required **threshold material gain** g_{th} to overcome the losses of the medium α_i and the mirror α_m. This gives $N_p = \frac{\eta_i(I-I_{th})}{qV_e v_g g_{th}}$. The desired power P_{out} escaping a mirror is

$$(N_pV_p)\cdot\hbar\omega\cdot\underbrace{v_g\alpha_m}_{1/\tau_m} = \underbrace{\eta_i\frac{\alpha_m}{\alpha_i+\alpha_m}}_{\eta_d}\cdot\frac{\hbar\omega}{q}(I - I_{th}) \implies \boxed{P_{out} = \eta_d\frac{\hbar\omega}{q}(I - I_{th}),} \qquad (29.28)$$

which relates the output power P_{out} of photon energy $\hbar\omega$ to the injected excess current $I - I_{th}$ via a **differential quantum efficiency** η_d as seen in Figures 29.22 (a) and (b). The output power exhibits a characteristic abrupt transition from spontaneous emission to primarily stimulated emission for $I > I_{th}$.

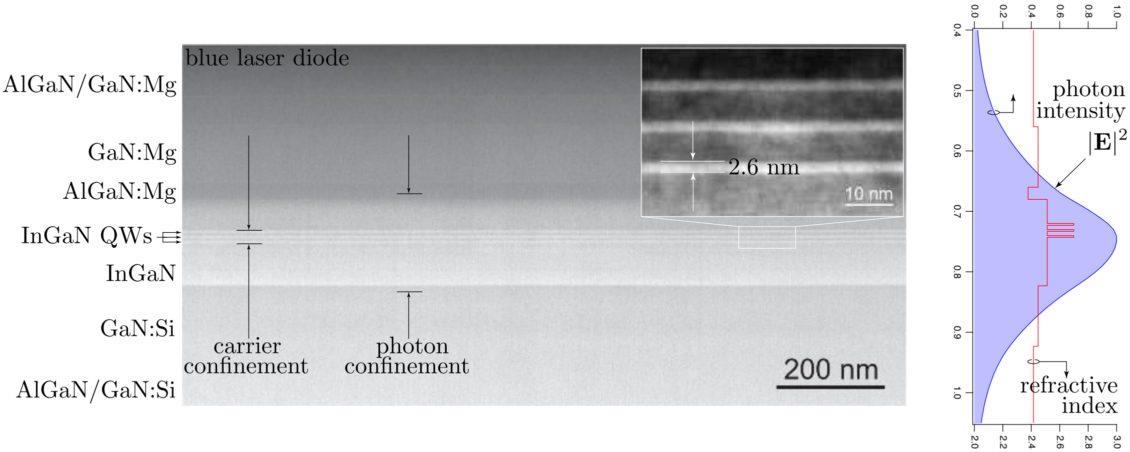


Fig. 29.23 A GaN based blue laser diode heterostructure with the refractive index profile and optical mode. The zoomed in image shows the three InGaN quantum wells which are the regions of optical gain. Kindly provided by Dr. Greg Muziol and Dr. Henryk Turski, IHPP PAS, UNIPRESS, Warsaw, Poland.

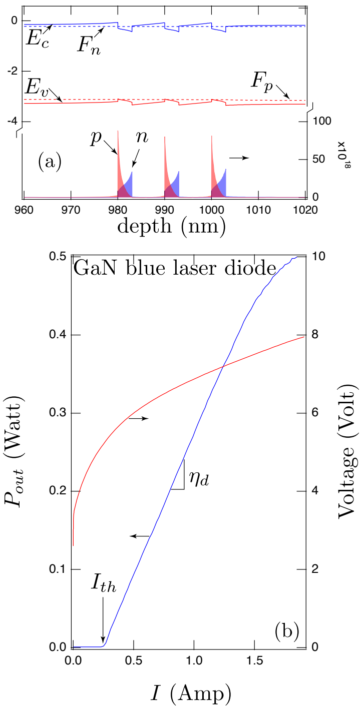


Fig. 29.24 (a) Energy band diagram (in eV units) and carrier concentration (in cm^{-3} units) above threshold, and (b) measured output characteristics of the GaN laser diode of Fig. 29.23. The threshold current I_{th} and differential efficiency η_d are indicated.

Assuming $V_e = Ad$, $I_{th}/A = J_{th} = qdN_{th}/(\tau\eta_i)$, which for $N_{th} \approx 10^{18}/\text{cm}^3$, $d \sim 100$ nm, $\eta_i \approx 1$ and $\tau \approx 10^{-9}$ s gives $J_{th} \approx 1600$ A/cm^2. Typical threshold current densities can be 5–10 × lower than this value depending on τ and other parameters (also see Exercise 29.4).

Large signal behavior: Fig. 29.22 (d) shows the large-signal turn-on behavior of a single-mode laser. If a constant current is injected starting at time $t = 0$, the output power "rings" before stabilizing. This is very similar to an *RLC* oscillator, and for good reason. The output power is related linearly to N_p, which is related to the injected carrier density N via Equations 29.27. Since the gain g depends on the injected carrier density N via the Fermi difference function which depends exponentially on the Fermi level, the gain $g \approx g_0 \ln(\frac{N}{N_{tr}})$ is an excellent approximation. Here g_0 is a characteristic material gain and N_{tr} is the **transparency carrier concentration** at which the gain becomes positive (but is still lower than the loss), as shown in Fig. 29.22 (c). Equations 29.27 (a) and (b) are coupled since N depends on N_p and N_p depends on N. A numerical solution yields the oscillations of the carrier concentration N and output power P_{out} of the kind shown in Fig. 29.22 (d), where the oscillations are typically in ns time scale.

Small signal behavior: If the laser diode is biased above threshold, and a (small) oscillating current $I(t) = I_0 + ie^{i\omega t}$ is applied on top of the DC bias, the carrier and photon densities respond as $N(t) = N_0 + ne^{i\omega t}$ and $N_p(t) = N_{p0} + n_p e^{i\omega t}$. Linearizing the gain as $g \approx a(N - N_{tr})$ where typical values of $a \sim \frac{g_0}{N_{tr}} \sim 10^{-16}$ cm^2, the response function obtained with these substitutions in Equations 29.27 is

$$G(\omega) = \frac{p(\omega)}{i(\omega)} = \frac{\eta_d \frac{\hbar\omega}{q}}{1 - (\frac{\omega}{\omega_R})^2 + i(\frac{\omega}{\omega_R})(\omega_R \tau_p + \frac{1}{\omega_R \tau})}, \tag{29.29}$$

where $\omega_R^2 = \frac{v_g a N_{p0}}{\tau_p} = \frac{\Gamma v_g a}{q V_e}(I - I_{th})$ is the square of the **relaxation resonance frequency**. The absolute squared AC response peaks at ω_R and drops off at 40 dB/decade at higher ω as shown in Fig. 29.22 (e). The cutoff frequencies increase with increasing current injection, and modern high-end laser diodes can follow inputs up to $\sim$ 100 GHz.

29.10 Case study: blue laser diode

Consider now a gallium nitride (GaN) based blue laser diode whose cross section is shown in Fig. 29.23. The gain medium consists of three $\sim$ 2.6 nm thick quantum wells of the alloy InGaN. The energy band diagram at a current injection that splits the quasi-Fermi levels to $F_c - F_v = 3.4$ eV is shown in Fig. 29.24 (a). The calculated injected electron and hole densities in the quantum wells range from $\sim 10^{18} - 10^{19}/\text{cm}^3$. The InGaN clad layers of lower In concentration (and hence larger energy bandgap) where there is no doping, extends $\sim$ 100 nm on both sides of the quantum wells. An electron blocking layer: Mg-doped AlGaN is inserted before the p-type GaN hole injection layer. The electron blocking layer is used to offset the rather large difference in electron and hole mobilities and diffusion constants in GaN.

The photon cavity of this diode laser is of the Fabry–Perot type, with cleaved facet mirrors, the same as in Fig. 29.18. The photon confinement layer is roughly $d = 200$ nm thick, nearly double of $\lambda_0/2n_r$ where $\lambda_0 \sim 450$ nm is the free space laser output wavelength of this device, and $n_r \sim 2.4$ is roughly the refractive index at that wavelength. The corresponding refractive index profile and the calculated optical mode profile $|\mathbf{E}|^2$ is shown in Fig. 29.23. It is clear that only a small portion $\Gamma = \int_{QWs} |\mathbf{E}|^2 dz / \int_{-\infty}^{+\infty} |\mathbf{E}|^2 dz$ of the cavity photons see the gain upon population inversion in the quantum wells. $\Gamma \sim 0.02$ is typical in such structures, implying only a 2% overlap. This requires the material gain to be high to achieve threshold criteria for lasing.

The measured output power vs. current and the current-voltage characteristics of this diode are shown in Fig. 29.24 (b). Compare this curve with Figures 29.22 (a) and (b). The threshold current is $I_{th} \sim 200$ mA. Calculation of the differential efficiency and several associated parameters is left as an exercise for the reader (see Exercise 29.5).

The output spectra of the blue laser diode are shown in Fig. 29.25 (a – c) for increasing injection current. Compare this with Fig. 29.19. Note the very significant narrowing of the spectra, and the far-field pattern of the laser beam shown in the inset of Fig. 29.25 (c). Fig. 29.25 (d) shows the zoomed-in spectra showing several modes are lasing, though one dominates. This is highlighted in the logarithmic scale plot of the same spectra in Fig. 29.25 (e). Such multi-mode lasing is typical of edge-emitting lasers. The next section discusses a method to make the laser single-mode and linewidth narrowing of that mode.

The modal gain spectra $g(\hbar\omega)$ of the blue laser diode are shown in Fig. 29.26 (a) for increasing injection current. The corresponding peak

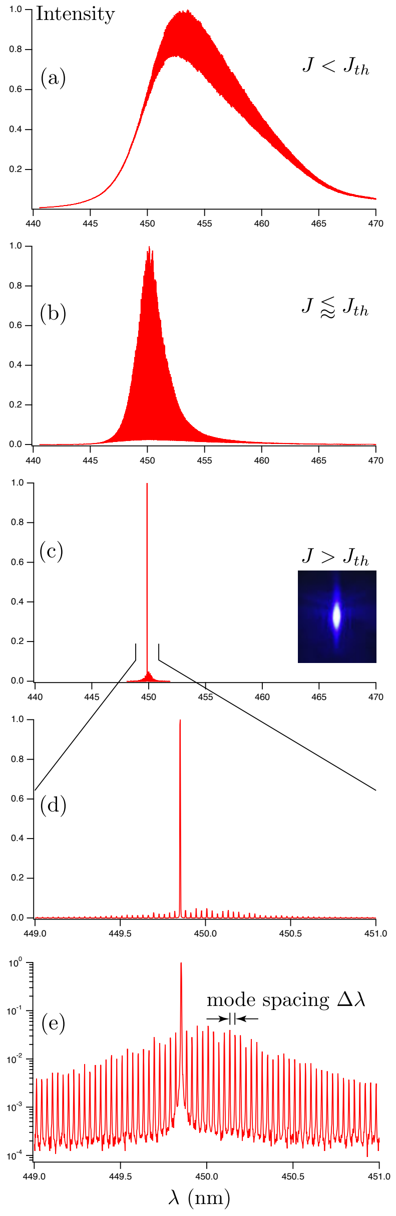


Fig. 29.25 Evolution of the edge-emitting blue laser diode output spectrum with increasing injection current density. (a) and (b) are the spectra for $J < J_{th}$, and (c) is for $J > J_{th}$. The laser beam is shown as an inset in (c), and (d) and (e) show the zoomed-in spectra in linear and log scales to highlight the cavity modes and their separation.

gain g_{max} as a function of the injection current density is shown in Fig. 29.26 (b). Compare Fig. 29.26 (a) with Fig. 29.19 (a), and Fig. 29.26 (b) with Fig. 29.22 (c). Several properties of the diode can be understood and appreciated with further analysis of the measured data presented in this section. Exercise 29.5 takes you through this analysis.

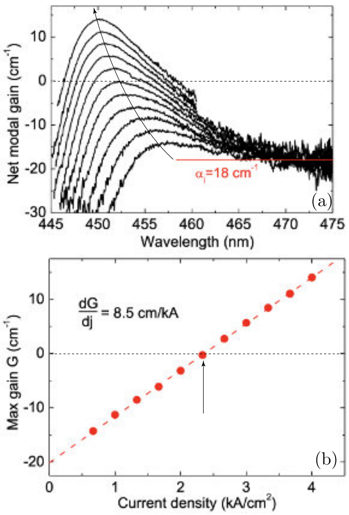
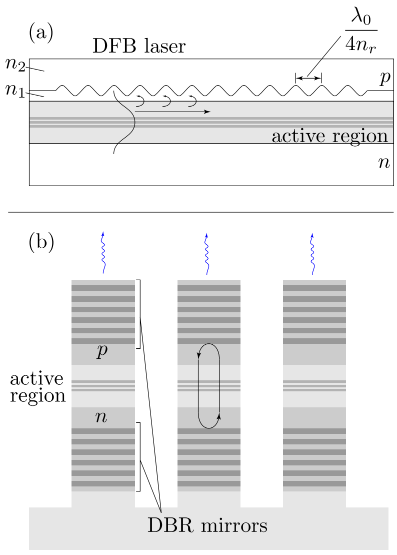


Fig. 29.26 (a) Net modal gain spectra at various injection current densities indicated by dots in (b) for the GaN blue laser diode of Fig. 29.23. The loss is 18 cm^{-1}. The maximum gain vs. the injection current density is shown in (b).

29.11 DFB, VCSELs and QCLs

Fig. 29.27 (a) shows the cross-section of a modified edge-emitting laser design. Here a periodic quarter-wavelength grating of spatial periodicity $\lambda_0/4n_r$ where λ_0 is the free-space laser emission wavelength and n_r is the average refractive index of the semiconductor is added. This is achieved by a combination of lithographic etching of the top layer of index n_1 and epitaxial regrowth of a layer of different refractive index n_2. This cavity is called a **distributed feedback Bragg** (DFB) grating, and they are called DFB lasers. A portion of the light wave propagating in the plane of the active region has a spatial overlap with this periodic grating. Because of multiple reflections, the grating allows only the mode of wavelength λ_0/n_r in the material (and λ_0 in free-space) to exist inside the cavity and causes destructive interference for other wavelengths. The single-wavelength emission leads to a **side-mode rejection ratio** (SMRR) of $\sim$ 30 dB or more. By modification of sections of the DFB grating, the effective n_r can be actively tuned by carrier injection, making such lasers also **electrically tunable** over a limited range of wavelengths. For the reasons of single-mode and tunable operation, DFB lasers are used for communication systems where closely spaced wavelengths are needed for data multiplexing.



Vertical Cavity Surface Emitting Laser (VCSEL)

Fig. 29.27 Various kinds of semiconductor lasers. (a) Distributed feedback Bragg (DFB) laser, and (b) vertical cavity surface emitting laser (VCSEL).

A major change in the output characteristics of lasers of all kinds, and especially of a single mode laser, is a significant narrowing of the output linewidth with increasing output power. The reason for this may be understood from the laser rate Equation 29.27 (b). The linewidth for the cavity without stimulated or spontaneous generation, obtained from $dN_p/dt = -N_p/\tau_p$ is simply (see Exercise 29.6) $\Delta\omega = 1/\tau_p$. For $I > I_{th}$ the rate equation written as $dN_p/dt = -(1/\tau_p - \Gamma v_g g)N_p + \Gamma\beta_{sp}R_{sp}$ already indicates a reduction in the rate of photon loss from the cavity, or in other words an increase in the photon lifetime in the cavity. This enhancement is the reason for the reduction of the linewidth to $\Delta\omega = 1/\tau_p - \Gamma v_g g = \frac{\Gamma\beta_{sp}R_{sp}}{N_p}$ in the steady state. For semiconductor laser diodes, accounting for the nonlinear coupling of the rate equations gives half of this value. This **modified Schawlow–Townes linewidth formula** is given by

$$\boxed{\Delta\omega = \frac{\Gamma\beta_{sp}R_{sp}}{2N_p} \propto \frac{1}{P_{out}},} \qquad (29.30)$$

which explicitly captures the reduction in the linewidth of the laser output spectrum with increasing output power.

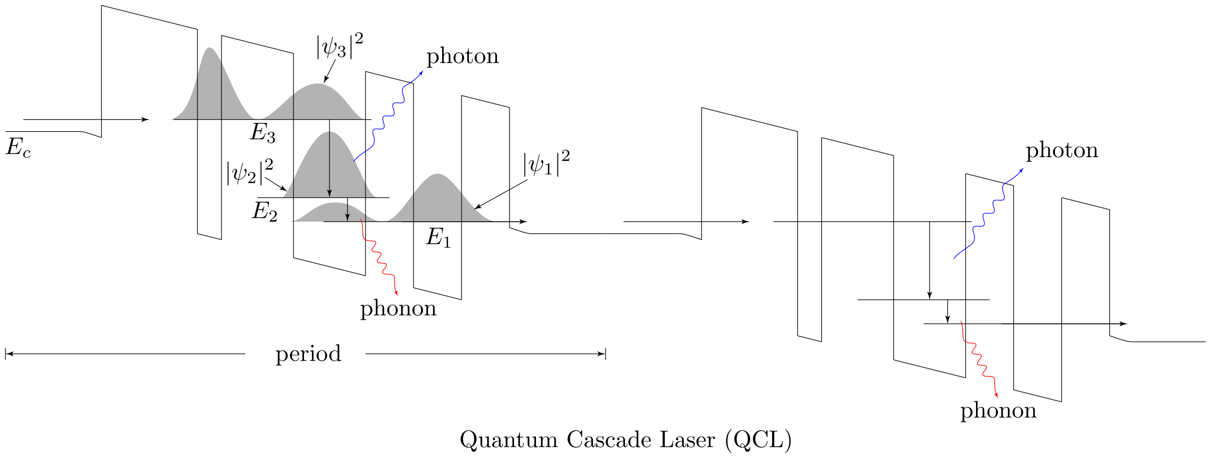


Fig. 29.28 Schematic energy band diagram of an intersubband quantum cascade laser (QCL).

Fig. 29.27 (b) shows the cross-section of a vertical cavity surface-emitting laser. The Bragg mirrors on top and bottom create a cavity in the vertical direction. The mirror reflectivities are much higher than cleaved facets, and can approach unity. A 2D array of VCSEL pillars can emit an array of laser beams. The Bragg mirrors are created either by epitaxial growth, or by deposition of dielectric stacks after growth of the active regions. The carrier injection can occur through the mirror, or through the sides utilizing lateral spreading. Though VCSELs are more difficult to fabricate compared to edge-emitting lasers and in general emit lower powers because of smaller cavities, the large array, nearly single-mode operation, and easy integration with vertical optical fibers or waveguides are distinct advantages for high bandwidth communication applications, and have widespread and increasing use in imaging, including for facial recognition in smartphones.

Fig. 29.28 shows the energy band diagram of two periods of a **quantum cascade laser** (QCL). Unlike the diode lasers, the QCL is a **unipolar** device that can operate with only electrons in the conduction band via intersubband transitions[16]. The QCL operates as if there is a three-level laser system in each period. Consider an electron injected from the left by resonant tunneling into the energy eigenvalue E_3 of the first thin well. The next well is thicker: the highest energy level lines up with the eigenvalue of the first well, and the electron effectively resonantly tunnels from the left into the excited state in the thick well. It takes a fraction of a ps to tunnel into state E_3 in the thicker well. When in this state, the electron can either relax to state E_2 non-radiatively by emission of phonons, or radiatively by the emission of a photon. The phonon emission time is several ps, while the intersubband photon emission time is much shorter. Therefore radiative recombination dominates, and simultaneously a population inversion is built up with larger electron population in E_3 than in E_2. The intersubband transi-

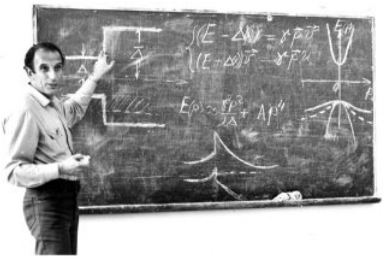


Fig. 29.29 The quantum cascade laser was first proposed by R. F. Kazarinov and R. A. Suris (pictured above) in the paper "Possibility of amplification of electromagnetic waves in a semiconductor with a superlattice," Fizika I Tekhnika Poluprovodnikov, 5 797 (1971).

[16] Interband cascade lasers have also been realized.

tion from $E_3 \to E_2$ produces a photon of energy $\hbar\omega = E_3 - E_2$. The electron then rapidly relaxes to state E_1 by phonon emission, and is pulled out to the right to the third well, ready to be injected to the next period, to repeat exactly the same process that happened in the first period. Thus, the electron *cascades* down the 3-level systems producing a photon in each $E_3 \to E_2$ transition. Because the occupation of state E_3 is larger than E_2 in the steady state, population inversion is achieved, which creates stimulated emission and optical gain[17].

The formation of cavities by cleaved facets, DFB, (or even external cavities), completes the construction of a quantum cascade laser. The QCL is a true marvel of quantum semiconductor heterostructure design and their epitaxial growth, and precision engineering of light-matter interaction in a condensed matter system. QCL output photon wavelengths can range from $\sim 2.6 - 250$ μm with corresponding energies of $\sim 480 - 5$ meV. QCLs typically require cooling for operation. For the longest wavelength QCLs, which operate at ~ 10 THz frequencies or above, the optical phonon energy of the semiconductor crystal can become a bottleneck since they are essential for rapid depopulation of the lower energy level to maintain population inversion.

QCLs are used for spectroscopic detection of environmental pollutants since the photon energies are resonant with the vibrational energies of molecules such as NO_2 or CO_2. Since the highest frequency semiconductor transistors can reach ~ 1 THz, and QCLs operate at above ~ 10 THz, there still exists a THz gap between $1 - 10$ THz to fill for high-efficiency semiconductor based compact power sources for a variety of applications in sensing, medicine, astrophysics, and other emerging areas.

29.12 Towards field quantization

A single-mode semiconductor laser with $P_{out} = 10$ mW at $\hbar\omega = 1$ eV emits $\sim \frac{P_{out}}{\hbar\omega} = 6.25 \times 10^{16}$ photons per second into a single optical mode. Similarly a LED emits enormous numbers of photons into a large number of modes. The number of modes are very large for blackbody radiation. For the rest of the chapter, we explore the **lower limit** of photon numbers in electromagnetic fields (i. .e., the quantum limit). In doing so, we will learn the core concepts of **quantum electrodynamics**, a fitting way to summarize the quantum concepts of this book, and to provide closure to deeper unanswered questions such as: a) why could we not compute the spontaneous emission rate directly using Fermi's golden rule earlier, and b) why did we give electrons (and matter) the quantum treatment but not consider the quantum nature of light itself? We will not be shy in introducing quantitative aspects of field theories here, because beyond the facade of mathematical notations, it is conceptually rather simple, and deeply satisfying in how it explains unanswered questions, and enriches what we have learned earlier. It also sets you up for the increasingly important con-

[17] Several options other than the three-quantum well (QW) design of Fig. 29.28 exist for the active region of the QCL, ranging from one QW, two QWs, to superlattices. The optical transitions are designed to be either from bound-to-bound states, bound-to-continuum states, inter-miniband states, and even interband states.

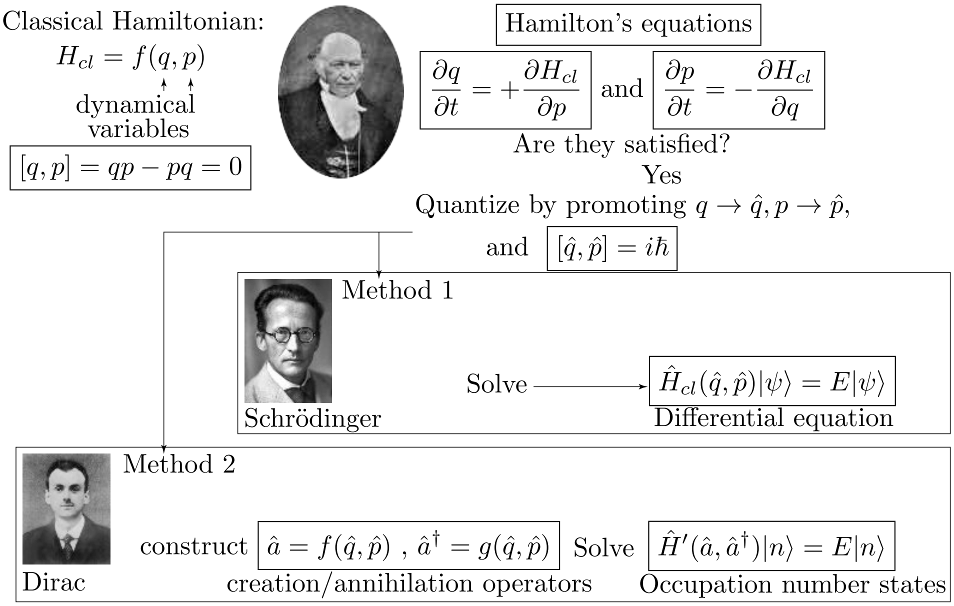

Classical Hamiltonian:
$H_{cl} = f(q,p)$
dynamical variables
$[q,p] = qp - pq = 0$
Hamilton's equations
$\frac{\partial q}{\partial t} = +\frac{\partial H_{cl}}{\partial p}$ and $\frac{\partial p}{\partial t} = -\frac{\partial H_{cl}}{\partial q}$
Are they satisfied?
Yes
Quantize by promoting $q \rightarrow \hat{q}, p \rightarrow \hat{p}$,
and $[\hat{q},\hat{p}] = i\hbar$
Method 1
Solve
$\hat{H}_{cl}(\hat{q},\hat{p})|\psi\rangle = E|\psi\rangle$
Differential equation
Schrödinger
Method 2
construct
$\hat{a} = f(\hat{q},\hat{p})$, $\hat{a}^\dagger = g(\hat{q},\hat{p})$
Solve
$\hat{H}'(\hat{a},\hat{a}^\dagger)|n\rangle = E|n\rangle$
creation/annihilation operators
Occupation number states
Dirac

Fig. 29.30 Quantization of a classical Hamiltonian after Hamilton's equations following Schrödinger and Dirac methods.

cepts of circuit and cavity quantum electrodynamics.

In Chapters 26 and 27 we treated the electromagnetic field as a classical object to compute the absorption coefficient of semiconductors in Chapter 27, and to understand photonic devices in Chapters 28 and 29. The electromagnetic field entered the Schrödinger equation via the momentum term $\mathbf{p} \rightarrow \mathbf{p} + q\mathbf{A}$. We found that a monochromatic wave of frequency $\omega = c|\mathbf{k}|$ is composed of an electric field $\mathbf{E} = \hat{\mathbf{e}}E_0 \sin(\mathbf{k}\cdot\mathbf{r} - \omega_{\mathbf{k}}t)$ and $\mathbf{A} = \hat{\mathbf{e}}A_0 \cos(\mathbf{k}\cdot\mathbf{r} - \omega_{\mathbf{k}}t)$ is the magnetic vector potential. The fields oscillate in the direction $\hat{\mathbf{e}}$ and linked by the dispersion $\omega_{\mathbf{k}} = c|\mathbf{k}|$, and $\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t}$. The classical energy stored in the field in a volume Ω for that mode is the **continuous** function

$$H_{em}(\omega) = \Omega \cdot [\frac{1}{2}\epsilon_0 E^2 + \frac{1}{2}\epsilon_0 \omega^2 A^2]. \qquad (29.31)$$

But the photoelectric effect and blackbody radiation require photon energies to be **discrete**. How can we then represent the discrete nature of photons in electromagnetic fields? Is there a "Schrödinger equation" for light?

How do we "quantize" a classical problem? If the classical energy is $H_{cl}(x,p)$, i.e., a function of the dynamical variables (x,p), then they satisfy Hamilton's equations[18] if the conditions $\frac{dx}{dt} = +\frac{\partial H_{cl}}{\partial p}$ and $\frac{dp}{dt} = -\frac{\partial H_{cl}}{\partial x}$ are met. We transition to quantum mechanics by *postulating* that $[\hat{x},\hat{p}] = i\hbar \neq 0$, and promoting the dynamical variables to operators. The Schrödinger equation $\hat{H}_{cl}(\hat{x},\hat{p})|\psi\rangle = E|\psi\rangle$ then becomes a *differential* equation that provides a recipe to find the quantum

[18] We encountered the Hamiltonian approach to classical mechanics in Chapter 21 when discussing the Liouville theorem in the lead-up to the Boltzmann equation.

states $|\psi\rangle$ and the allowed energies E. This is highlighted in Fig. 29.30.

We shall now use the all-important 1D harmonic oscillator problem to verify this route of quantization[19]. A particle of mass m experiences a potential $V(x) = \frac{1}{2}m\omega^2 x^2$ classically. The particle (see Fig. 29.31) then can have *any* positive energy $E = H_{cl}(x, p_x) = \frac{p_x^2}{2m} + \frac{1}{2}m\omega^2 x^2$, i.e., $0 \leq E \leq \infty$. The dynamical variables x, p_x are seen to satisfy Hamilton's equations $\frac{dx}{dt} = \frac{p_x}{m} = \frac{\partial H_{cl}}{\partial p_x}$ and $\frac{dp_x}{dt} = -m\omega^2 x = -\frac{\partial H_{cl}}{\partial x}$, and $[x, p_x] = 0$. To transition to quantum mechanics, we *postulate* that $[\hat{x}, \hat{p}_x] = i\hbar$. By solving the Schrödinger equation $[\frac{\hat{p}_x^2}{2m} + \frac{1}{2}m\omega^2\hat{x}^2]|\psi\rangle = E|\psi\rangle$, which is a differential equation, we find that the allowed energies E become restricted to $E_n = (n + \frac{1}{2})\hbar\omega$, where $n = 0, 1, 2, ...$. This was discussed in the exactly solvable problems in quantum mechanics in Chapter 5. The particle is forbidden from being at the bottom of the well because $E_0 = \frac{\hbar\omega}{2}$, the energy due to zero-point motion required to satisfy the uncertainty principle.

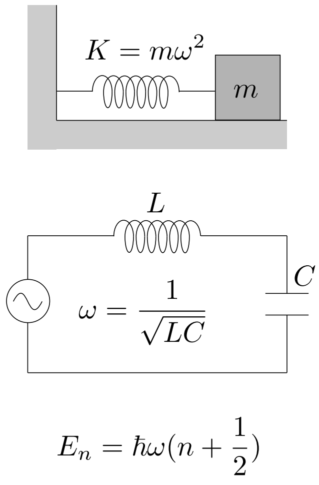

$K = m\omega^2$
m
L
C
$\omega = \frac{1}{\sqrt{LC}}$
$E_n = \hbar\omega(n + \frac{1}{2})$

Fig. 29.31 The mechanical mass-spring system and the electrical LC oscillator are both harmonic oscillator problems in quantum mechanics.

But photons have no mass. So at first glance the differential equation version of the Schrödinger equation cannot be simply carried over for quantizing the electromagnetic field! How can one then write a corresponding quantum equation for photons? The crucial breakthrough was made by Dirac, who realized that the harmonic oscillator problem itself need not be solved using the differential equation version of Schrödinger, but by an entirely different technique. The key to the technique is taking **square roots**! Note that if a, b are numbers, $a^2 + b^2 = (a - ib)(a + ib)$. So for the classical energy, we can write

$$H_{cl} = \frac{p_x^2}{2m} + \frac{1}{2}m\omega^2 x^2 = \hbar\omega[\sqrt{\frac{m\omega}{2\hbar}}(x - \frac{i}{m\omega}p_x)][\sqrt{\frac{m\omega}{2\hbar}}(x + \frac{i}{m\omega}p_x)]. \tag{29.32}$$

Note that the terms in the square brackets are dimensionless, and are complex conjugates of each other. This relation is an identity as long as x and p_x are numbers. Does it remain so if we transition to quantum mechanics and set $[\hat{x}, \hat{p}_x] = i\hbar$? Let us define the dimensionless *operators*

$$\boxed{\hat{a} = \sqrt{\frac{m\omega}{2\hbar}}(\hat{x} + \frac{i}{m\omega}\hat{p}_x)} \text{ and } \boxed{\hat{a}^\dagger = \sqrt{\frac{m\omega}{2\hbar}}(\hat{x} - \frac{i}{m\omega}\hat{p}_x),} \tag{29.33}$$

and the corresponding space and momentum operators in terms of them:

$$\boxed{\hat{x} = \sqrt{\frac{\hbar}{2m\omega}}(\hat{a}^\dagger + \hat{a})} \text{ and } \boxed{\hat{p}_x = i\sqrt{\frac{m\omega\hbar}{2}}(\hat{a}^\dagger - \hat{a}).} \tag{29.34}$$

The right hand side of the energy Equation 29.32 then reads $\hbar\omega\hat{a}^\dagger\hat{a}$. Using the commutator $[\hat{x}, \hat{p}_x] = i\hbar$, we find that

$$\hbar\omega\hat{a}^\dagger\hat{a} = \frac{m\omega^2}{2}[\hat{x}^2 + \frac{\hat{p}_x^2}{m^2\omega^2} + \frac{i}{m\omega}(\hat{x}\hat{p}_x - \hat{x}\hat{p}_x)] = [\frac{\hat{p}_x^2}{2m} + \frac{1}{2}m\omega^2\hat{x}^2] - \frac{\hbar\omega}{2}, \tag{29.35}$$

---

[19] The mechanical mass-spring oscillator is identical to the electrical LC circuit.

i.e., we are off by $-\hbar\omega/2$ because of the transition from numbers to operators. However, the order of the operators matter:

$$\hbar\omega\hat{a}\hat{a}^\dagger = \frac{m\omega^2}{2}[\hat{x}^2 + \frac{\hat{p}_x^2}{m^2\omega^2} - \frac{i}{m\omega}(\hat{x}\hat{p}_x - \hat{x}\hat{p}_x)] = [\frac{\hat{p}_x^2}{2m} + \frac{1}{2}m\omega^2\hat{x}^2] + \frac{\hbar\omega}{2}, \tag{29.36}$$

which means the commutator of the operators is (subtracting Equation 29.35 from 29.36)

$$\boxed{\hat{a}\hat{a}^\dagger - \hat{a}^\dagger\hat{a} = [\hat{a}, \hat{a}^\dagger] = 1.} \tag{29.37}$$

This also means that the harmonic oscillator Hamiltonian operator may be written completely in terms of the operators $\hat{a}$ and $\hat{a}^\dagger$:

$$\boxed{\hat{H}_{osc} = \frac{\hat{p}_x^2}{2m} + \frac{1}{2}m\omega^2\hat{x}^2 = \hbar\omega(\hat{a}^\dagger\hat{a} + \frac{1}{2}).} \tag{29.38}$$

Note that the mass m does not appear explicitly in the new form of the Hamiltonian: it is of course hiding "under the hood" of the operators $\hat{a}$ and $\hat{a}^\dagger$. Let us now investigate a few properties of the new operators. Is there a function ψ for which $\hat{a}\psi = 0$? Let's define $l = \sqrt{\hbar/m\omega}$ as the characteristic length scale of the harmonic oscillator problem. By direct substitution, we find $\langle x|\hat{a}|\psi\rangle = (x + \frac{\hbar}{m\omega}\frac{d}{dx})\psi = (x + l^2\frac{d}{dx})\psi = 0$ has the solution $\psi(x) = Ae^{-\frac{x^2}{2l^2}}$. Normalizing the function over $-\infty < x < +\infty$, we get $A = (\pi l)^{-1/4}$. This is a surprise, because the Schrödinger route had told us that the Gaussian $\langle x|0\rangle = \psi_0(x) = (\pi l)^{-1/4}e^{-x^2/2l^2}$ is exactly the eigenfunction of the ground state! So the operator $\hat{a}$ seems to be **destroying the ground state** of the harmonic oscillator, because $\langle x|\hat{a}|0\rangle = \hat{a}\langle x|0\rangle = \hat{a}\psi_0(x) = 0$. We also find that $\hat{a}^\dagger\psi_0 = \psi_1$, i.e., $\hat{a}^\dagger|0\rangle = |1\rangle$. It does not need much more work to show that in general the action of the operators on state $|n\rangle$ of the harmonic oscillator yields

$$\boxed{\hat{a}|n\rangle = \sqrt{n}|n-1\rangle}, \ \boxed{\hat{a}^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle}, \text{ and } \boxed{\hat{a}^\dagger\hat{a}|n\rangle = n|n\rangle}. \tag{29.39}$$

We realize then that $\hat{a}$ is an **annihilation operator**: it lowers the state of the system one rung of the harmonic oscillator ladder. It truncates at the lower end because $\hat{a}|0\rangle = 0$. This is shown in Fig. 29.32. Similarly, its Hermitian conjugate $\hat{a}^\dagger$ raises it one rung, and is a **creation operator**. For example, $\hat{a}^\dagger|9\rangle = \sqrt{10}|10\rangle$, and $\hat{a}|10\rangle = \sqrt{10}|9\rangle$. Together in the form $\hat{a}^\dagger\hat{a} = \hat{n}$, they form the **particle number operator**, meaning $\hat{a}^\dagger\hat{a}$ counts the quantum number of the state[20].

Equation 29.38 then gives the eigenvalues E_n when the Hamiltonian acts on eigenstates $|n\rangle$

$$\boxed{\hat{H}_{osc}|n\rangle = \hbar\omega(\hat{a}^\dagger\hat{a} + \frac{1}{2})|n\rangle = \underbrace{\hbar\omega(n + \frac{1}{2})}_{E_n}|n\rangle.} \tag{29.40}$$

Let us then recap Dirac's alternate quantum theory of the harmonic oscillator problem. If the dynamical variables (v, w) of a classical problem with energy $H_{cl}(v, w)$ satisfy Hamilton's equations $\frac{dv}{dt} = u = \frac{\partial H_{cl}}{\partial w}$

[20] For example, $\hat{a}^\dagger\hat{a}|10\rangle = \hat{a}^\dagger(\sqrt{10}|9\rangle) = 10|10\rangle$. Also note that we can create any state we want by starting from the ground state $|0\rangle$ and repeatedly acting on it by the creation operator: $|n\rangle = \frac{(\hat{a}^\dagger)^n}{\sqrt{n!}}|0\rangle$. For example, $|3\rangle = \frac{(\hat{a}^\dagger)^3}{\sqrt{3.2.1}}|0\rangle = \frac{(\hat{a}^\dagger)^2}{\sqrt{3.2.1}}(\sqrt{1})|1\rangle = \frac{(\hat{a}^\dagger)}{\sqrt{3.2.1}}(\sqrt{2.1})|2\rangle = \frac{1}{\sqrt{3.2.1}}(\sqrt{3.2.1})|3\rangle$.

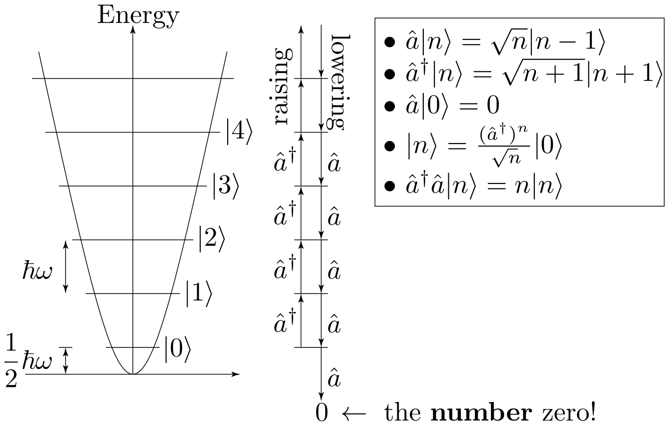

Energy
$|4\rangle$
$|3\rangle$
$|2\rangle$
$|1\rangle$
$|0\rangle$
$\hbar\omega$
$\frac{1}{2}\hbar\omega$
raising
lowering
$\hat{a}^\dagger$
$\hat{a}$
• $\hat{a}|n\rangle = \sqrt{n}|n-1\rangle$
• $\hat{a}^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle$
• $\hat{a}|0\rangle = 0$
• $|n\rangle = \frac{(\hat{a}^\dagger)^n}{\sqrt{n}}|0\rangle$
• $\hat{a}^\dagger\hat{a}|n\rangle = n|n\rangle$
0 ← the **number** zero!

Fig. 29.32 Actions of the creation and annihilation operators.

and $\frac{dw}{dt} = -\omega^2 v = -\frac{\partial H_{cl}}{\partial u}$, then we can quantize the problem by postulating $[\hat{v}, \hat{w}] = i\hbar \neq 0$. We can then find the corresponding creation and annihilation operators $\hat{a}^\dagger$ and $\hat{a}$ as linear combinations of $\hat{v}$ and $\hat{w}$. The allowed states $|\psi\rangle$ in the quantum picture are then those that satisfy the Hamiltonian $\hbar\omega(\hat{a}^\dagger\hat{a} + \frac{1}{2})|\psi\rangle = E|\psi\rangle$, subject to the operator relations in Equations 29.39. The potential $\frac{1}{2}m\omega^2x^2$ and kinetic $\frac{p_x^2}{2m}$ components of energy have "vanished" into a natural energy scale $\hbar\omega$ and the dimensionless creation and annihilation operators. The problem is now solved in its entirety *without* recourse to $\hat{x}$ and $\hat{p}_x$ with their classical interpretations, but by the use of $\hat{a}$ and $\hat{a}^\dagger$, with the much simplified interpretation of raising or lowering.

If we consider the zero-point energy $\hbar\omega/2$ as a reference, then the new energy $E' = \hbar\omega\hat{a}^\dagger\hat{a} = \hbar\omega\hat{n}$ scheme has a delightful alternate interpretation. The allowed eigenenergies form an equidistant energy ladder $E' = n\hbar\omega$. One particle in state $|n\rangle$ has the same energy as n particles in state $|1\rangle$. The action of a creation operator then indeed looks like creating a new particle of the same energy, and the annihilation operator seems to remove one particle. This form of quantizing the "field" is called **second quantization**. It is the beginning of the canonical formulation of quantum field theory.

We are now ready to quantize the electromagnetic field using Dirac's insight. Going back to the energy content of a monochromatic electromagnetic field in volume Ω:

$$H_{em}(\omega) = \frac{1}{2}\Omega\epsilon_0 E^2 + \frac{1}{2}\Omega\epsilon_0\omega^2 A^2, \tag{29.41}$$

we note the uncanny similarity to the classical harmonic oscillator problem. If we associate the electric field to the momentum via $p_x \leftrightarrow -\epsilon_0\Omega E$, the vector potential to space coordinate via $x \leftrightarrow A$, and the quantity $\epsilon_0\Omega$ to the mass of the oscillator via $m \leftrightarrow \epsilon_0\Omega$, then the electromagnetic field energy transforms exactly into $\frac{p_x^2}{2m} + \frac{1}{2}m\omega^2x^2$, the harmonic oscillator energy. They become mathematically identical. We should now follow Dirac's prescription and check if the classical vector field pair $\mathbf{E}$ and $\mathbf{A}$ satisfy Hamilton's equations, and then proceed to the quantization of the field if they do. To do that, let us first generalize from a monochromatic wave to a general **broadband wave**[21]. If the mathematics appears involved at any point, it is useful to go back to the monochromatic picture. But the broadband treatment of the electromagnetic field will enable us to obtain far more insights to the problem.

29.13 Broadband field modes

From Faraday's law $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$ and the definition of the vector potential $\mathbf{B} = \nabla \times \mathbf{A}$, we realize that $\nabla \times (\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}) = 0$. But since $\nabla \times (\nabla \phi) = 0$ for any scalar function ϕ, we have in general $\mathbf{E} = -\nabla \phi - \frac{\partial \mathbf{A}}{\partial t}$. We consider an electromagnetic wave propagating in free space with no free charge. Gauss's law requires $\nabla \cdot \mathbf{E} = 0$. Then,

$$\nabla^2 \phi + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = 0 \tag{29.42}$$

must be satisfied. This is where we need to choose a gauge. To see why, one can shift the scalar potential background by a constant scalar $\phi \to \phi + \phi_0$ and still satisfy Equation 29.42. Furthermore, if we choose a scalar function χ and form the vector field $\nabla \chi$, then $\mathbf{B} = \nabla \times (\mathbf{A} + \nabla \chi) = \nabla \times \mathbf{A}$, because $\nabla \times (\nabla \chi) = 0$ is a vector identity. It will prove useful to choose the scalar field χ to ensure that $\nabla \cdot \mathbf{A} = 0$. This choice of gauge is called the Coulomb gauge[22]. From Equation 29.42, we should then have $\nabla^2 \phi = 0$. We also choose $\phi = 0$. This is the gauge we choose to work with. With this choice, Maxwell's equations gives $\mathbf{B} = \mu_0 \mathbf{H} = \nabla \times \mathbf{A}$, and $\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t}$ *uniquely*. From Ampere's law, we get $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = -\frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2}$. Since $\nabla \cdot \mathbf{A} = 0$, we get the wave equation

$$\nabla^2 \mathbf{A} = \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2}. \tag{29.43}$$

Solutions of this Maxwell wave equation yields the allowed vector potentials $\mathbf{A}(\mathbf{r}, t)$. From there, we obtain the electric field $\mathbf{E}(\mathbf{r}, t) = -\frac{\partial \mathbf{A}(\mathbf{r},t)}{\partial t}$ and $\mathbf{H}(\mathbf{r}, t) = \frac{1}{\mu_0}\nabla \times \mathbf{A}(\mathbf{r}, t)$ *uniquely* because the gauge has been fixed. The classical problem is then completely solved if we find the allowed solutions for $\mathbf{A}(\mathbf{r}, t)$.

Recall the procedure for solving the time-dependent Schrödinger equation[23] $\hat{H}\psi = i\hbar\frac{\partial \psi}{\partial t}$, which is of a similar nature as Equation 29.43 (albeit with $\frac{\partial(...)}{\partial t}$ on the RHS). We found the subset of solutions that allowed a *separation* of the time and space variables were *eigenstates*: they formed a *complete* set of mutually orthogonal *modes*. Let us follow the same prescription here, and search for solutions of the vector potential of the form $\mathbf{A}(\mathbf{r}, t) = \mathbf{A}(\mathbf{r}) \cdot T(t)$. Realizing that Equation 29.43 is a compact version of three independent equations because

---

[21] Blackbody radiation or the output of LEDs are broadband in which multiple modes are occupied by photons.

[22] As mentioned earlier in the text, nobody ever reads a paper in which someone has done an experiment involving photons with the footnote that says "This experiment was done in Coulomb gauge". This quote from Sidney Coleman highlights that the *physics* is independent of the gauge choice. It is chosen for mathematical convenience for a particular problem. For example, for electrons in a constant magnetic field, the Landau gauge $\mathbf{A} = (0, Bx, 0)$ is often used. In electrodynamics, the Lorenz gauge $\nabla \cdot \mathbf{A} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0$ is also popular.

[23] From Chapter 3, Section 3.8.

$\mathbf{A}(\mathbf{r}) = \hat{\mathbf{x}}A_x + \hat{\mathbf{y}}A_y + \hat{\mathbf{z}}A_z$, we find

$$T(t)\nabla^2 A_x = \frac{1}{c^2}A_x\frac{d^2T(t)}{dt^2} \implies$$
$$c^2\frac{\nabla^2 A_x}{A_x} = \frac{1}{T(t)}\frac{d^2T(t)}{dt^2} = \text{constant} = -\omega^2 \tag{29.44}$$

must hold for all three spatial components (x, y, z). The x-dependent and the t-dependent LHS terms depend on separate variables. So they can be equal only if they are both equal to the same constant. Setting the constant to be $-\omega^2$, we find solutions of the time (temporal) part are of the form $T(t) = T(0)e^{\pm i\omega t}$. For the spatial part, we realize that functions of the form $A_x(x, y, z) = A_x(0, 0, 0)e^{\pm i(k_x\cdot x+k_y\cdot y+k_z\cdot z)} = A_x(0, 0, 0)e^{\pm i\mathbf{k}\cdot\mathbf{r}}$ satisfy $c^2\frac{\nabla^2 A_x}{A_x} = -\omega^2$ if

$$\boxed{\omega = \omega_{\mathbf{k}} = c\sqrt{k_x^2 + k_y^2 + k_z^2} = c|\mathbf{k}|.} \tag{29.45}$$

From all the allowed solutions, the subset that allow the separation of time and space variables are of the form $\mathbf{A}(\mathbf{r}, t) = \hat{\mathbf{e}}A_{\mathbf{k}}e^{\pm i\mathbf{k}\cdot\mathbf{r}}e^{\pm i\omega_{\mathbf{k}}t}$ where $\hat{\mathbf{e}}$ is the unit vector along $\mathbf{A}(\mathbf{r}, t)$, $\omega_{\mathbf{k}} = c|\mathbf{k}|$, and $A_{\mathbf{k}}$ is a *number* characterizing the strength of the vector potential, i.e., it is *not* a function of $\mathbf{r}$. Since $\nabla \cdot \mathbf{A} = i(\mathbf{k}\cdot\hat{\mathbf{e}})A_{\mathbf{k}}e^{\pm i\mathbf{k}\cdot\mathbf{r}}e^{\pm i\omega_{\mathbf{k}}t} = 0$, $\hat{\mathbf{e}}\cdot\mathbf{k} = 0$, implying we must have $\hat{\mathbf{e}} \perp \mathbf{k}$, where $\mathbf{k}$ is the direction of propagation of the wave. It represents a linearly polarized TEM wave. If $\mathbf{k} = k\hat{\mathbf{z}}$, then $\hat{\mathbf{e}} = \hat{x}$ and $\hat{\mathbf{e}} = \hat{y}$ are independent solutions that satisfy $\hat{\mathbf{e}}\cdot\mathbf{k} = 0$. They represent physically different waves. Thus, **two polarizations** are allowed. Let us label the polarization as s, with the understanding that $\hat{\mathbf{e}}_{s_1} = \hat{\mathbf{x}}$ and $\hat{\mathbf{e}}_{s_2} = \hat{\mathbf{y}}$ for the wave here. The two polarization states are orthogonal, because $\hat{\mathbf{e}}_{s_1}\cdot\hat{\mathbf{e}}_{s_2} = 0$. The allowed solutions may then be written as $\mathbf{A}(\mathbf{r}, t) = \sum_{s,\mathbf{k}}\hat{\mathbf{e}}_s A_{\mathbf{k}}e^{\pm i\mathbf{k}\cdot\mathbf{r}}e^{\pm i\omega_{\mathbf{k}}t}$. It is evident that this is a Fourier decomposition.

The sum over $\mathbf{k}$ should really be an integral if *all* values of $\mathbf{k}$ are allowed. But every integral is the limit of a sum. Here, it is both more physical and convenient to work with discrete $\mathbf{k}$ values, and then move to the continuum. To see why, imagine placing a cubic box of side L with the origin at $\mathbf{r} = (0, 0, 0)$ as shown in Fig. 29.33, and requiring the vector potential to meet periodic boundary conditions[24]. Then, $\mathbf{A}(x + L, y, z, t) = \mathbf{A}(x, y, z, t)$ for each allowed $\mathbf{k}$ requires $e^{ik_xL} = 1$, restricting it to $\mathbf{k} = (k_x, k_y, k_z) = \frac{2\pi}{L}(n_x, n_y, n_z)$, where $n_x = 0, \pm1, \pm2, ...$ can only take integer values. The set of allowed $\mathbf{k}$ vectors then form a 3D discrete lattice as shown in Fig. 29.33. Each lattice point represents a *mode* of the wave. For each mode (n_x, n_y, n_z), there are two allowed polarizations s_1, s_2. We can write the mode index compactly by defining $\lambda = (s, n_x, n_y, n_z)$, with the understanding $-\lambda = (s, -n_x, -n_y, -n_z)$. In that case, we have $\mathbf{k}_{-\lambda} = -\mathbf{k}_\lambda$ and $\omega_{-\lambda} = \omega_\lambda$. This just means that the wave indexed by $-\lambda$ is moving in the opposite direction to that indexed by λ with the same polarization and same frequency.

---

[24] This is not a restriction, because one can take $L \to \infty$ in the end. But it is physical, because if indeed there was a cubic cavity such as in a laser, it would enforce the fields to go to *zero* at the mirrors.

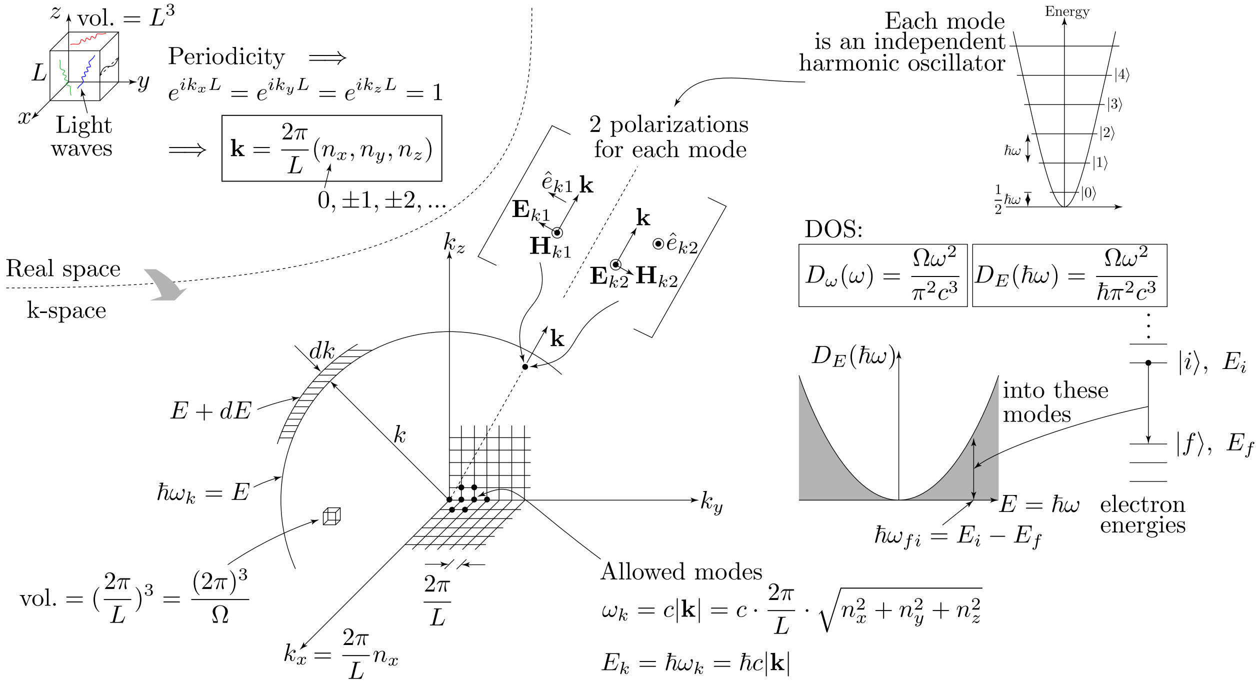

z
vol. $= L^3$
L
y
x
Light waves
Periodicity $\implies$
$e^{ik_xL} = e^{ik_yL} = e^{ik_zL} = 1$
$\implies \boxed{\mathbf{k} = \frac{2\pi}{L}(n_x, n_y, n_z)}$
$0, \pm1, \pm2, \ldots$
Real space
k-space
2 polarizations for each mode
$\hat{e}_{k1}$ $\mathbf{k}$
$\mathbf{E}_{k1}$
$\mathbf{H}_{k1}$
$\mathbf{k}$
$\hat{e}_{k2}$
$\mathbf{E}_{k2}$ $\mathbf{H}_{k2}$
Each mode is an independent harmonic oscillator
Energy
$|4\rangle$
$|3\rangle$
$|2\rangle$
$|1\rangle$
$|0\rangle$
$\hbar\omega$
$\frac{1}{2}\hbar\omega$
DOS:
$D_\omega(\omega) = \frac{\Omega\omega^2}{\pi^2 c^3}$
$D_E(\hbar\omega) = \frac{\Omega\omega^2}{\hbar\pi^2 c^3}$
k_z
$\mathbf{k}$
dk
$E + dE$
k
$\hbar\omega_k = E$
k_y
vol. $= (\frac{2\pi}{L})^3 = \frac{(2\pi)^3}{\Omega}$
$\frac{2\pi}{L}$
$k_x = \frac{2\pi}{L} n_x$
Allowed modes
$\omega_k = c|\mathbf{k}| = c \cdot \frac{2\pi}{L} \cdot \sqrt{n_x^2 + n_y^2 + n_z^2}$
$E_k = \hbar\omega_k = \hbar c|\mathbf{k}|$
$D_E(\hbar\omega)$
into these modes
$|i\rangle,\ E_i$
$|f\rangle,\ E_f$
$E = \hbar\omega$
electron energies
$\hbar\omega_{fi} = E_i - E_f$

Fig. 29.33 Allowed modes for electromagnetic waves.

Now we can ask the following question: how many modes are available between frequencies ω and $\omega + d\omega$, or between energies E and $E + dE$? You have already seen this in Chapters 3, 4, and 5. This number is the density of states (DOS) of photons. From Fig. 29.33, since the dispersion relation is $\omega = c|\mathbf{k}|$, the modes of constant energy lie on the surface of a sphere defined by $E = \hbar\omega = \hbar c|\mathbf{k}|$. Modes $\omega + d\omega$ and corresponding energies $E + dE$ form a slightly bigger sphere. The thin spherical shell between the two spheres has a volume $4\pi k^2 dk$ in the 3D k-space. How many modes fall inside this shell? Now because the volume occupied by each mode in the $\mathbf{k}$-space is $(\frac{2\pi}{L})^3 = \frac{(2\pi)^3}{\Omega}$, and because each mode allows two polarizations, the DOS is given by

$$D_\omega(\omega)d\omega = D_E(\hbar\omega)d(\hbar\omega) = 2 \cdot \frac{4\pi k^2 dk}{\frac{(2\pi)^3}{\Omega}}$$

$$\implies \boxed{D_\omega(\omega) = \frac{\omega^2\Omega}{\pi^2 c^3}} \text{ and } \boxed{D_E(\hbar\omega) = \frac{\omega^2\Omega}{\hbar\pi^2 c^3}}. \qquad (29.46)$$

The photon DOS in 3D increases as the *square* of the frequency (or energy) as shown in Fig. 29.33. What this means is as we increase the energy, there are more modes that have the same energy. The DOS depends on the dimensionality of the problem, we have considered three-dimensions here.

Because the modes indexed by λ form a complete set, the most general *real* solutions of the wave Equation 29.43 are

$$\mathbf{A}(\mathbf{r},t) = \sum_{\lambda}[\hat{\mathbf{e}}_{\lambda}A_{\lambda}e^{i(\mathbf{k}_{\lambda}\cdot\mathbf{r}-\omega_{\lambda}t)} + \hat{\mathbf{e}}_{\lambda}A_{\lambda}^{\star}e^{-i(\mathbf{k}_{\lambda}\cdot\mathbf{r}-\omega_{\lambda}t)}], \tag{29.47}$$

where the second term is the complex conjugate of the first. Note that if $\hat{\mathbf{e}}_{\lambda} = \hat{\mathbf{x}}$ and $A_{\lambda} = A_0/2$ for a mode, the vector potential of that mode is $\mathbf{A}_{\lambda}(\mathbf{r},t) = \hat{\mathbf{x}}A_0\cos(\mathbf{k}_{\lambda}\cdot\mathbf{r}-\omega_{\lambda}t)$. Because we will need to find the energy content of the wave by finding $\mathbf{E}$ and $\mathbf{H}$ from $\mathbf{A}$, we choose to club together the terms in the expansion in the following fashion. Merge the vector and spatial part by defining the mode vector $\mathbf{M}_{\lambda}(\mathbf{r}) = \hat{\mathbf{e}}_{\lambda}e^{i\mathbf{k}_{\lambda}\cdot\mathbf{r}}$. Merge the scalar field strength and the time dependence of the mode into the scalar function $Q_{\lambda}(t) = A_{\lambda}e^{-i\omega_{\lambda}t}$. With these definitions, Equation 29.47 is rewritten as

$$\mathbf{A}(\mathbf{r},t) = \sum_{\lambda}[Q_{\lambda}(t)\mathbf{M}_{\lambda}(\mathbf{r}) + Q_{\lambda}^{\star}(t)\mathbf{M}_{\lambda}^{\star}(\mathbf{r})]. \tag{29.48}$$

We now claim that in the above sum in Equation 29.48, the mode vectors $\mathbf{M}_{\lambda}(\mathbf{r})$ corresponding to different λ are orthogonal. To see why this is true, evaluate

$$\begin{aligned}\int d^3r\mathbf{M}_{\lambda}(\mathbf{r})\cdot\mathbf{M}_{\nu}^{\star}(\mathbf{r}) &= \\ (\hat{\mathbf{e}}_{\lambda}\cdot\hat{\mathbf{e}}_{\nu})\int_0^L e^{i(k_{x\lambda}-k_{x\nu})x}dx\int_0^L e^{i(k_{y\lambda}-k_{y\nu})y}dy\int_0^L e^{i(k_{z\lambda}-k_{z\nu})z}dz.\end{aligned} \tag{29.49}$$

Because $k_{x\lambda} = \frac{2\pi}{L}n_{x\lambda}...$, the integrals $\int_0^L dxe^{i\frac{2\pi}{L}(n_{x\lambda}-n_{x\nu})x} = L\delta_{n_{x\lambda},n_{x\nu}}$ lead to

$$\int d^3r\mathbf{M}_{\lambda}(\mathbf{r})\cdot\mathbf{M}_{\nu}^{\star}(\mathbf{r}) = L^3\delta_{\lambda,\nu} = \Omega\delta_{\lambda,\nu}, \tag{29.50}$$

where $\Omega = L^3$ is the volume of the cube. This proves they are orthogonal. Now let us find $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{H}(\mathbf{r},t)$ from the vector potential in Equation 29.48. Since $\frac{\partial Q_{\lambda}(t)}{\partial t} = -i\omega_{\lambda}Q_{\lambda}(t)$, we obtain

$$\mathbf{E}(\mathbf{r},t) = -\frac{\partial\mathbf{A}(\mathbf{r},t)}{\partial t} = \sum_{\lambda}\underbrace{i\omega_{\lambda}[Q_{\lambda}(t)\mathbf{M}_{\lambda}(\mathbf{r}) - Q_{\lambda}^{\star}(t)\mathbf{M}_{\lambda}^{\star}(\mathbf{r})]}_{\mathbf{E}_{\lambda}}, \tag{29.51}$$

and since $\nabla\times\mathbf{M}_{\lambda}(\mathbf{r}) = i\mathbf{k}_{\lambda}\times\mathbf{M}_{\lambda}(\mathbf{r})$, we get

$$\mathbf{H}(\mathbf{r},t) = \frac{1}{\mu_0}\nabla\times\mathbf{A}(\mathbf{r},t) = \frac{i}{\mu_0}\sum_{\lambda}\underbrace{\mathbf{k}_{\lambda}\times[Q_{\lambda}(t)\mathbf{M}_{\lambda}(\mathbf{r}) - Q_{\lambda}^{\star}(t)\mathbf{M}_{\lambda}^{\star}(\mathbf{r})]}_{\mathbf{H}_{\lambda}}. \tag{29.52}$$

The total *classical* energy of the electromagnetic wave in the cube is given by

$$H_{em} = \int d^3r[\frac{1}{2}\epsilon_0\mathbf{E}\cdot\mathbf{E} + \frac{1}{2}\mu_0\mathbf{H}\cdot\mathbf{H}] \implies$$

$$\boxed{H_{em} = \epsilon_0\Omega\sum_{\lambda}\omega_{\lambda}^2[Q_{\lambda}(t)Q_{\lambda}^{\star}(t) + Q_{\lambda}^{\star}(t)Q_{\lambda}(t)].} \tag{29.53}$$

Showing this requires some algebra, but the fact that it must be independent of $\mathbf{M}_\lambda(\mathbf{r})$ could have been anticipated because this is the total energy in the cubic box. Further, we have written $2Q_\lambda(t)Q^\star_\lambda(t) = Q_\lambda(t)Q^\star_\lambda(t) + Q^\star_\lambda(t)Q_\lambda(t)$ in a *symmetric* form in anticipation of it appearing in the quantized Hamiltonian, where it will necessary need to be Hermitian. To make the connection to the harmonic oscillator problem more explicit, motivated by Dirac's method of taking "square roots" in Equation 29.32, we define

$$\boxed{x_\lambda(t) = \sqrt{\epsilon_0\Omega}[Q_\lambda(t) + Q^\star_\lambda(t)]}$$

$$\text{and } \boxed{p_\lambda(t) = -i\omega_\lambda\sqrt{\epsilon_0\Omega}[Q_\lambda(t) - Q^\star_\lambda(t)],} \tag{29.54}$$

and the inverse relations

$$\boxed{Q_\lambda(t) = \frac{1}{\sqrt{4\epsilon_0\Omega}}[x_\lambda(t) + \frac{i}{\omega_\lambda}p_\lambda(t)]}$$

$$\text{and } \boxed{Q^\star_\lambda(t) = \frac{1}{\sqrt{4\epsilon_0\Omega}}[x_\lambda(t) - \frac{i}{\omega_\lambda}p_\lambda(t)].} \tag{29.55}$$

With these definitions, the *classical* electromagnetic energy in the cubic box becomes

$$\boxed{H_{em} = \sum_\lambda H_\lambda = \sum_\lambda[\frac{1}{2}p^2_\lambda(t) + \frac{1}{2}\omega^2_\lambda x^2_\lambda(t)],} \tag{29.56}$$

where it is clear that the energy is distributed over various modes λ and each term in the sum in Equation 29.56 is the energy content of that mode. This is what we mean by **broadband**. All energies including zero are allowed. Each term in the sum is equivalent to the energy content of a harmonic oscillator with mass $m = 1$, for which $x_\lambda(t) = x_\lambda(0)\cos(\omega_\lambda t)$ and $p_\lambda(t) = \frac{d}{dt}x_\lambda(t) = -x_\lambda(0)\omega_\lambda\sin(\omega_\lambda t)$, and $H_{\lambda,osc} = \frac{1}{2}\omega^2_\lambda x^2_\lambda(0)$. Equation 29.56 is now ready for quantization.

29.14 Quantization of fields

Following Dirac's prescription, we identify $x_\lambda(t)$ and $p_\lambda(t)$ as the dynamical variables of mode λ from the classical energy in Equation 29.56. Do they satisfy Hamilton's equations? Because $Q_\lambda(t) = A_\lambda e^{-i\omega_\lambda t}$, we find

$$\frac{dx_\lambda(t)}{dt} = \sqrt{\epsilon_0\Omega}\frac{d}{dt}[Q_\lambda(t) + Q^\star_\lambda(t)]$$

$$= -i\sqrt{\epsilon_0\Omega\omega^2_\lambda}[Q_\lambda(t) - Q^\star_\lambda(t)] = p_\lambda(t) = \frac{\partial H_\lambda}{\partial p_\lambda}, \tag{29.57}$$

implying the first Hamilton's equation is indeed satisfied. Similarly,

$$\begin{aligned}\frac{dp_\lambda(t)}{dt} &= -i\sqrt{\epsilon_0\Omega\omega_\lambda^2}\frac{d}{dt}[Q_\lambda(t) - Q_\lambda^\star(t)] \\ &= \omega_\lambda^2\sqrt{\epsilon_0\Omega}[Q_\lambda(t) + Q_\lambda^\star(t)] = -\omega_\lambda^2 x_\lambda(t) = -\frac{\partial H_\lambda}{\partial x_\lambda} \qquad (29.58)\end{aligned}$$

is also satisfied. Thus, $x_\lambda(t)$ and $p_\lambda(t)$ form a classically conjugate pair that commute. Dirac's prescription is then to promote them from scalars to Hermitian operators, and enforce the quantization condition $[\hat{x}_\lambda(t), \hat{p}_\lambda(t)] = \hat{x}_\lambda(t)\hat{p}_\lambda(t) - \hat{p}_\lambda(t)\hat{x}_\lambda(t) = i\hbar$. Because of the orthogonality of different modes λ, we can write this compactly as $[\hat{x}_\lambda(t), \hat{p}_\nu(t)] = i\hbar\delta_{\lambda,\nu}$, and collect the ancillary relations $[\hat{x}_\lambda(t), \hat{x}_\nu(t)] = 0$ and $[\hat{p}_\lambda(t), \hat{p}_\nu(t)] = 0$. Note that the operators born out of promoting the dynamical variables are explicitly time-dependent. They are in the **Heisenberg picture** of quantum mechanics. In the **Schrödinger picture** of quantum mechanics, quantum state vectors $|\psi(t)\rangle$ are time-dependent and operators $\hat{p}_\lambda$ are time-independent. In the Heisenberg picture the state vectors $|\psi\rangle$ are time-independent, all the time-evolution is in the operators $\hat{p}_\lambda(t)$. We identify the creation and annihilation operators for mode λ in analogy to Equation 29.33.

Using the classical definitions in Equations 29.54 and 29.55 we write the quantum version of the annihilation operator as

$$\begin{aligned}\hat{a}_\lambda(t) &= \sqrt{\frac{\omega_\lambda}{2\hbar}}[\hat{x}_\lambda(t) + \frac{i}{\omega_\lambda}\hat{p}_\lambda(t)] \\ &= \sqrt{\frac{2\epsilon_0\Omega\omega_\lambda}{\hbar}}\hat{Q}_\lambda(t) = \sqrt{\frac{2\epsilon_0\Omega\omega_\lambda}{\hbar}}\hat{A}_\lambda e^{-i\omega_\lambda t}. \qquad (29.59)\end{aligned}$$

The classically scalar quantity $\hat{A}_\lambda$ has now become an operator. It had the physical meaning of the strength of the vector potential in the classical version, and will retain this meaning in the quantum version as the operator whose **expectation value** is the strength of the vector potential. Also note the explicit time-dependence of the annihilation operator in the Heisenberg representation is simply $\hat{a}_\lambda(t) = \hat{a}_\lambda(0)e^{-i\omega_\lambda t}$.

Let us denote $\hat{a}_\lambda = \hat{a}_\lambda(0) = \sqrt{\frac{2\epsilon_0\Omega\omega_\lambda}{\hbar}}\hat{A}_\lambda$. Following the same procedure, we write the creation operator as

$$\begin{aligned}\hat{a}_\lambda^\dagger(t) &= \sqrt{\frac{\omega_\lambda}{2\hbar}}[\hat{x}_\lambda(t) - \frac{i}{\omega_\lambda}\hat{p}_\lambda(t)] \\ &= \sqrt{\frac{2\epsilon_0\Omega\omega_\lambda}{\hbar}}\hat{Q}_\lambda^\dagger(t) = \sqrt{\frac{2\epsilon_0\Omega\omega_\lambda}{\hbar}}\hat{A}_\lambda^\dagger e^{+i\omega_\lambda t}. \qquad (29.60)\end{aligned}$$

Denote the time-dependence of the operator by $\hat{a}_\lambda^\dagger(t) = \hat{a}_\lambda^\dagger(0)e^{+i\omega_\lambda t} = \hat{a}_\lambda^\dagger e^{+i\omega_\lambda t}$, where $\hat{a}_\lambda^\dagger = \sqrt{\frac{2\epsilon_0\Omega\omega_\lambda}{\hbar}}\hat{A}_\lambda^\dagger$. We also have the relations

$$\hat{x}_\lambda(t) = \sqrt{\epsilon_0\Omega}[\hat{Q}_\lambda(t) + \hat{Q}_\lambda^\dagger(t)] = \sqrt{\frac{\hbar}{2\omega_\lambda}}[\hat{a}_\lambda^\dagger(t) + \hat{a}_\lambda(t)], \qquad (29.61)$$

$$\hat{p}_\lambda(t) = -i\omega_\lambda\sqrt{\epsilon_0\Omega}[\hat{Q}_\lambda(t) - \hat{Q}^\dagger_\lambda(t)] = i\sqrt{\frac{\hbar\omega_\lambda}{2}}[\hat{a}^\dagger_\lambda(t) - \hat{a}_\lambda(t)]. \qquad (29.62)$$

Now since $[\hat{x}_\lambda(t), \hat{p}_\lambda(t)] = i\hbar$, we get

$$[\sqrt{\frac{\hbar}{2\omega_\lambda}}(\hat{a}^\dagger_\lambda(t) + \hat{a}_\lambda(t)), i\sqrt{\frac{\hbar\omega_\lambda}{2}}(\hat{a}^\dagger_\lambda(t) - \hat{a}_\lambda(t))] = i\hbar$$

$$\implies [\hat{a}^\dagger_\lambda(t) + \hat{a}_\lambda(t), \hat{a}^\dagger_\lambda(t) - \hat{a}_\lambda(t)] = 2 \qquad (29.63)$$

from where we obtain the commutator for the creation and annihilation operators for each mode λ to be

$$\boxed{[\hat{a}_\lambda(t), \hat{a}^\dagger_\lambda(t)] = 1.} \qquad (29.64)$$

Substituting $\hat{Q}_\lambda(t) = \sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}\hat{a}_\lambda(t)$ and $\hat{Q}^\dagger_\lambda(t) = \sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}\hat{a}^\dagger_\lambda(t)$ into the classical energy Equation 29.53 we get

$$\begin{aligned}\hat{H}_{em} &= \epsilon_0\Omega\sum_\lambda \omega^2_\lambda[\hat{Q}_\lambda(t)\hat{Q}^\dagger_\lambda(t) + \hat{Q}^\dagger_\lambda(t)\hat{Q}_\lambda(t)] \\ &= \sum_\lambda \frac{1}{2}\hbar\omega_\lambda[\hat{a}_\lambda(t)\hat{a}^\dagger_\lambda(t) + \hat{a}^\dagger_\lambda(t)\hat{a}_\lambda(t)], \qquad (29.65)\end{aligned}$$

and with the use of the commutator in Equation 29.64, it becomes

$$\boxed{\hat{H}_{em} = \sum_\lambda \hbar\omega_\lambda[\hat{a}^\dagger_\lambda(t)\hat{a}_\lambda(t) + \frac{1}{2}] = \sum_\lambda \hbar\omega_\lambda[\hat{a}^\dagger_\lambda\hat{a}_\lambda + \frac{1}{2}].} \qquad (29.66)$$

We have used $\hat{a}^\dagger_\lambda(t)\hat{a}_\lambda(t) = \hat{a}^\dagger_\lambda e^{+i\omega_\lambda t}\hat{a}_\lambda e^{-i\omega_\lambda t} = \hat{a}^\dagger_\lambda\hat{a}_\lambda$, which makes it clear that the net energy is indeed time-independent. Equation 29.66 completes the quantization of the broadband electromagnetic field. It is telling us that each mode of the field indexed by λ acts as an **independent harmonic oscillator** (Fig. 29.34). The total energy is the sum of the energy of each mode. We collect the expressions for the relevant operators and fields in this quantized version before discussing the rich physics that will emerge from the quantization. The vector potential from Equation 29.48 is now written as the operator

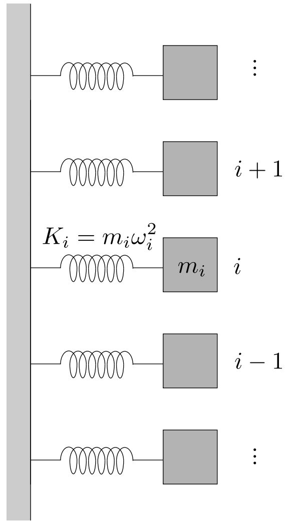


Fig. 29.34 Schematic representation of independent harmonic oscillators.

$$\boxed{\hat{\mathbf{A}}(\mathbf{r}, t) = \sum_\lambda \sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}\hat{\mathbf{e}}_\lambda[\hat{a}_\lambda e^{i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)} + \hat{a}^\dagger_\lambda e^{-i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)}],} \qquad (29.67)$$

the electric field operator is

$$\boxed{\hat{\mathbf{E}}(\mathbf{r}, t) = i\sum_\lambda \sqrt{\frac{\hbar\omega_\lambda}{2\epsilon_0\Omega}}\hat{\mathbf{e}}_\lambda[\hat{a}_\lambda e^{i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)} - \hat{a}^\dagger_\lambda e^{-i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)}],} \qquad (29.68)$$

and the magnetic field operator is

$$\boxed{\hat{\mathbf{H}}(\mathbf{r}, t) = \frac{i}{\mu_0}\sum_\lambda \sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}(\mathbf{k}_\lambda \times \hat{\mathbf{e}}_\lambda)[\hat{a}_\lambda e^{i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)} - \hat{a}^\dagger_\lambda e^{-i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)}].} \qquad (29.69)$$

If we want to revert from the Heisenberg to the Schrödinger picture for the time-dependent fields $\hat{\mathbf{H}}(\mathbf{r}, t)$ or $\hat{\mathbf{E}}(\mathbf{r}, t)$, we simply set $t = 0$ in Equations [25] 29.67 and 29.68. The creation and annihilation operators are dimensionless. Note the nature of the "field operators": the space and time dependence takes the nature of a wave, whereas **the creation and annihilation operators are independent of the space and time and act only on the occupation number of the modes**. Let us now investigate the repercussions of quantization of the EMag field.

29.15 Field quantization: aftermath

The physics of quantized field operators: Because the creation and annihilation operators do not commute, it is clear that $\hat{\mathbf{E}}$ and $\hat{\mathbf{H}}$ will now fail to commute for a single mode. This means **one cannot simultaneously measure the electric and magnetic fields of a mode of electromagnetic wave with arbitrary accuracy**. Even when there are absolutely no photons in the electromagnetic field, the net energy is infinite, as is seen from Equation 29.66. This remains an unresolved problem in quantum field theory (that is, if it is considered to be a problem!). The zero-point vibrations of electric and magnetic fields in the vacuum trigger spontaneous emission. One school of thought is that the electromagnetic vacuum is composed of particle-antiparticle pairs (such as electrons and positrons) and indeed stores infinite energy. The **Casimir effect** where two metal plates in close proximity attract each other is explained from this picture of vacuum fluctuations, though there exist alternate explanations for this phenomenon.

Occupation number formalism, Fock states: We have dutifully followed Dirac's method for quantizing the broadband electromagnetic field, culminating in the quantized Hamiltonian field operator in Equation 29.66. What then are the corresponding allowed eigenstates and eigenvalues of the field? They must satisfy the equation $\hat{H}_{em}|\psi\rangle = E|\psi\rangle$. Recall the quantization of the **single** harmonic oscillator in Equations 29.38 and 29.40, which said that if the oscillator is in the eigenstate $|n\rangle$, its energy is $E_n = \hbar\omega(n + \frac{1}{2})$. The broadband field Hamiltonian in Equation 29.66 is clearly a sum of the Hamiltonians of **independent** modes, where each of them is an oscillator. The energies simply add, meaning the eigenstate must be composed of a **product** of the independent eigenstates of each mode. Let us label an allowed eigenstate of mode λ by $|n_\lambda\rangle$, and form the composite eigenstate $|\psi\rangle = |n_{\lambda_1}\rangle|n_{\lambda_2}\rangle|n_{\lambda_2}\rangle$.... Since the Hamiltonian has the number operator $\hat{H}_{em} = \sum_\lambda \hbar\omega_\lambda(\hat{n}_\lambda + \frac{1}{2})$, when it acts on the constructed eigenstate, because $\hat{n}_{\lambda 1}|n_{\lambda 1}\rangle = n_{\lambda 1}|n_{\lambda 1}\rangle$ it will give us the total energy E_{tot}:

---

[25] Verify that each of the four expressions in Equations 29.66, 29.67, 29.68, and 29.69 are dimensionally correct.

$$\hat{H}_{em}|\psi\rangle = [\sum_\lambda \hbar\omega_\lambda(\hat{n}_\lambda + \frac{1}{2})]|n_{\lambda_1}\rangle|n_{\lambda_2}\rangle|n_{\lambda_2}\rangle ... =$$

$$\underbrace{[\sum_\lambda \hbar\omega_\lambda(n_\lambda + \frac{1}{2})]}_{E_{tot}=\sum_\lambda E_{n_\lambda}} |n_{\lambda_1}\rangle|n_{\lambda_2}\rangle|n_{\lambda_2}\rangle ... \qquad (29.70)$$

This implies each mode of the broadband field can only change its energy by *discrete* amounts $E_{n_\lambda+1} - E_{n_\lambda} = \hbar\omega_\lambda$. **We call the "particle" that carries this discrete packet of energy the photon**. It is very important to note that the **frequency** of a mode $\omega_\lambda = c|\mathbf{k}| = \frac{2\pi c}{L}\sqrt{n_x^2 + n_y^2 + n_y^2}$ is not really quantized, and could well be continuous for an infinite volume as $\Omega = L^3 \to \infty$. **What really is quantized is the *number* of photons n_λ in the mode λ.**

Fig. 29.35 Vladimir Fock made significant contributions to quantum mechanics and introduced the occupation number formalism.

The importance of the number of photons in each mode motivates an alternate way of writing the eigenstate: we can simply write the composite eigenstate as $|\psi\rangle = |n_{\lambda 1}, n_{\lambda 2}, n_{\lambda 3}, ...\rangle$, where $n_{\lambda 1}$ is the number of photons in mode λ_1, $n_{\lambda 2}$ is the number of photons in mode λ_2, and so on. This way of writing the state vector is called the **occupation number representation**, and the state is called a **Fock state** after the physicist Vladimir Fock (Fig. 29.35).

Since there are an infinite number of modes, the Fock space is infinitely large. Each of the photon numbers can assume the integer values $n_\lambda = 0, 1, 2, ...$ For example, $|0_1, 5_2, 30_3, 0_4, ...\rangle$ is an allowed state with 0 photons in mode 1, 5 photons in mode 2, and so on. The state $|0\rangle = |0_1, 0_2, 0_3, ...\rangle$ is a perfectly respectable state: it represents a quantum state of the electromagnetic field when there are no photons in *any* mode. This unique state is called the **vacuum state**.

The action of the photon creation and annihilation operators are now

$$\hat{a}_\lambda|n_1, n_2, ..., n_\lambda, ...\rangle = \sqrt{n_\lambda}|n_1, n_2, ..., n_\lambda - 1, ...\rangle, \text{ [raising]}$$
$$\hat{a}^\dagger_\lambda|n_1, n_2, ..., n_\lambda, ...\rangle = \sqrt{n_\lambda + 1}|n_1, n_2, ..., n_\lambda + 1, ...\rangle \text{ [lowering]}. \qquad (29.71)$$

So we have $\hat{a}^\dagger_\lambda|.., 5_\lambda, ...\rangle = \sqrt{6}|.., 6_\lambda, ...\rangle$. Similarly, we have $\hat{a}_\lambda|0\rangle = 0$. This says that the annihilation operator kills the vacuum state, and converts it to the *number* 0. The number 0 is *not* the same as the vacuum state $|0\rangle$! Because of the orthogonality of different modes, we have the inner product between allowed eigenstates

$$\langle m_1, m_2, ..., m_\nu, ...|n_1, n_2, ..., n_\lambda, ...\rangle = \delta_{n_1,m_1}\delta_{n_2,m_2}...\delta_{n_\lambda,m_\nu}..., \qquad (29.72)$$

which suggests that *any* quantum state of the broadband electromagnetic field may be written as the linear combination of the complete set of eigenstates

$$|\psi\rangle = [\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} ...]C_{n_1,n_2,...}|n_1, n_2, ...\rangle. \qquad (29.73)$$

Here $|C_{n_1,n_2,...}|^2$ is the probability of finding the field in a state with n_1 photons in mode 1, n_2 in mode 2, and so on. The allowed number of eigenstates is vast, and the Fock-state is enormous, larger than the Hilbert space. Much of this space is unexplored. Let us now look at a few states of interest.

"Thermal" photons: The Bose–Einstein distribution: From Equation 29.73, consider a state $|\psi\rangle = |1_1, 3_2, 0_3, ...\rangle$. This is a *correlated* state of the photon field, because it requires 1 photon in mode 1, and 3 photons in mode 2 **simultanesouly**. Consider the special case of a highly **uncorrelated** photon field state which looks like $|\psi\rangle_{thermal} = (\sum_{n_1=0}^{\infty} C_{n_1}|n_1, 0_2, 0_3, ...\rangle) + (\sum_{n_2=0}^{\infty} C_{n_2}|0_1, n_2, 0_3, ...\rangle) + ...$ Taking the inner product, we see the reason it is called uncorrelated: $\langle\psi|\psi\rangle_{thermal} = \sum_{n_1=0}^{\infty} |C_{n_1}|^2 + \sum_{n_2=0}^{\infty} |C_{n_2}|^2 + ...$, meaning there are specific probabilities $|C_{n\lambda}|^2$ of finding the state with n_λ photons in mode λ, and *no photons* in other modes. The probability of finding photons in any one mode is completely *independent* of how many photons there are in other modes. The probability of finding 10 photons in the third mode would be $|C_{n_3=10}|^2$.

This particular uncorrelated state of the photon field is a very important one – because it gave birth to quantum mechanics. It is physically realized for "thermal" photons, when the energy in the photon modes is in equilibrium with a thermal bath characterized by a temperature T. Boltzmann statistics state that if photons could be in many energy states E_m, then the probability of occupation of state E_n is $P(E_n) = e^{-\frac{E_n}{k_b T}}/Z$, where $Z = \sum_m e^{-\frac{E_m}{k_b T}}$ is the **partition function**, and k_b is the Boltzmann constant. Clearly $\sum_n P(E_n) = 1 = \langle\psi|\psi\rangle$. Consider a mode of the field with energy spacing $\hbar\omega_\lambda$. We learned from field quantization in Section 29.14 that the allowed energy values corresponding to n_λ photons occupying that mode is $E_{n_\lambda} = \hbar\omega_\lambda(n_\lambda + \frac{1}{2})$. The probability of n_λ photons being present in that mode is then

$$P(E_{n_\lambda}) = \frac{e^{-\frac{\hbar\omega_\lambda}{k_b T}(n_\lambda + \frac{1}{2})}}{\sum_{m_\lambda=0}^{\infty} e^{-\frac{\hbar\omega_\lambda}{k_b T}(m_\lambda + \frac{1}{2})}} = e^{-n_\lambda \frac{\hbar\omega_\lambda}{k_b T}}(1 - e^{-\frac{\hbar\omega_\lambda}{k_b T}}), \tag{29.74}$$

and the *average* number of photons in that mode is

$$\langle n_\lambda\rangle = \sum_{n_\lambda=0}^{\infty} n_\lambda P(E_{n_\lambda}) = \sum_{n_\lambda=0}^{\infty} n_\lambda e^{-n_\lambda \frac{\hbar\omega_\lambda}{k_b T}}(1 - e^{-\frac{\hbar\omega_\lambda}{k_b T}})$$

$$\implies \boxed{\langle n_\lambda\rangle = \frac{1}{e^{\frac{\hbar\omega_\lambda}{k_b T}} - 1}.} \tag{29.75}$$

We have thus arrived at a very fundamental result: that the statistical thermal average number of photons $\langle n_\lambda\rangle = \langle\hat{a}_\lambda^\dagger \hat{a}_\lambda\rangle_{thermal}$ in mode ω_λ in incoherent light in equilibrium with a bath at temperature T is given by the boxed Equation 29.75. This is the Bose–Einstein distribution function, the root of which can be tracked all the way back

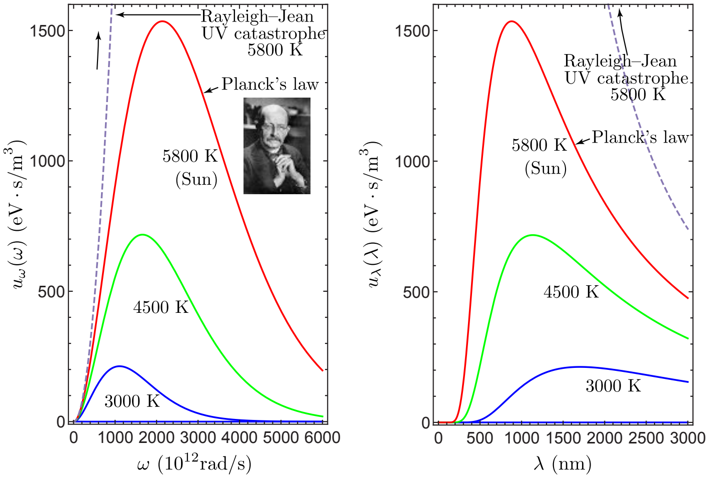


Fig. 29.36 Planck's law resolved the ultraviolet catastrophe.

to the fact that the particle (photon) creation and annihilation operators follow the relation $[\hat{a}_\lambda, \hat{a}^\dagger_\lambda] = 1$ which we got in Equation 29.64. *Any* bosonic particle that follows this commutation relation will have a corresponding Bose–Einstein distribution.

Planck's law of blackbody radiation: Now, we can ask the question: what is the *energy content* in the thermal photon field between modes ω_λ and $\omega_\lambda + d\omega_\lambda$? The mean energy in the mode is $\langle n_\lambda \rangle \hbar\omega_\lambda$, and the number of such modes is given by the photon DOS, $D_E(\hbar\omega_\lambda)$ from Equation 29.46. The energy *density* per unit volume is then, skipping the subscript λ,

$$U_\omega(\omega)d\omega = D_E(\hbar\omega)d(\hbar\omega) \cdot \langle n \rangle \hbar\omega$$

$$\implies \frac{U_\omega(\omega)}{\Omega} = \boxed{u_\omega(\omega) = \frac{\hbar\omega^3}{\pi^2 c^3} \cdot \frac{1}{e^{\frac{\hbar\omega}{k_b T}} - 1}.} \qquad (29.76)$$

What we did here is Max Planck's derivation of the blackbody radiation spectrum. The boxed part of Equation 29.76 is Planck's law for the blackbody spectrum, plotted in Fig. 29.36 for various temperatures as a function of frequency and wavelength. Also shown is the pre-quantum result of Rayleigh–Jeans, which is $u^{rj}_\omega(\omega) \approx \frac{\omega^2}{\pi^2 c^3} k_b T$, which diverges for high frequencies (the ultraviolet catastrophe). The Rayleigh–Jeans law is the high-temperature limit of Planck's law, and is easily derived assuming $k_b T >> \hbar\omega$ in Equation 29.76. Planck's constant does not appear in it. Planck's hypothesis that light came in quantized packets of energy $\hbar\omega$ resolved the ultraviolet catastrophe and gave birth to quantum mechanics.

What may not be clear now is: where does the temperature T come from? Maxwell's equations do not have temperature anywhere, so clearly the light is in equilibrium with some *other* objects? Yes, those other objects are composed of matter – specifically atoms and crystals that have electrons. Einstein re-derived Planck's law of radiation by invoking **thermodynamic equilibrium** between matter and light, mediated by the fundamental light-matter interaction processes. In the next section, we will consider this problem and show that it also holds the secrets of *spontaneous emission*, something that is remarkably difficult to explain without the quantization of the electromagnetic field we have achieved in this chapter. This requires us to tackle the light-matter interaction problem in its full quantum glory, as we do in the next section.

29.16 Quantized light-matter interaction

The classical Hamiltonian for a particle of mass M and charge Q in the presence of electromagnetic fields is

$$H_{\text{cl}} = \frac{(\mathbf{p} - Q\mathbf{A})^2}{2M} + V(\mathbf{r}) + \int d^3r [\frac{1}{2}\epsilon_0 \mathbf{E} \cdot \mathbf{E} + \frac{1}{2}\mu_0 \mathbf{H} \cdot \mathbf{H}], \qquad (29.77)$$

where $\mathbf{p}$ is the kinetic momentum of the particle, $\mathbf{A}$ is the magnetic vector potential of the electromagnetic field, $V(\mathbf{r})$ is the potential in which the particle is moving, $\mathbf{E}$ is the electric field and $\mathbf{H}$ is the magnetic field. We recognize the integral term as the energy content of the electromagnetic wave alone. We will consider the interaction of light primarily with electrons ($M = m_e, Q = -q$ where $q = +1.6 \times 10^{-19}$ C is the magnitude of the electron charge) and neglect the interaction with protons in the nucleus. One justification is their large mass, which makes their contribution to the Hamiltonian small[26]. The quantum version of the light-matter Hamiltonian then takes the form

$$\hat{H}_{\text{tot}} = \frac{(\hat{\mathbf{p}} + q\hat{\mathbf{A}})^2}{2m_e} + V(\mathbf{r}) + \int d^3r [\frac{1}{2}\epsilon_0 \hat{\mathbf{E}} \cdot \hat{\mathbf{E}} + \frac{1}{2}\mu_0 \hat{\mathbf{H}} \cdot \hat{\mathbf{H}}], \qquad (29.78)$$

where the hats represent operators. We are now treating electrons and photons at equal footing, by quantizing both. From our treatment of field quantization we use Equation 29.66 for the field energy to write

$$\hat{H}_{\text{tot}} = \frac{(\hat{\mathbf{p}} + q\hat{\mathbf{A}})^2}{2m_e} + V(\mathbf{r}) + \sum_{\lambda} \hbar\omega_\lambda (\hat{a}^\dagger_\lambda \hat{a}_\lambda + \frac{1}{2}). \qquad (29.79)$$

Squaring the first term (while keeping the operator order intact) and rearranging, we get

$$\boxed{\hat{H}_{\text{tot}} = \underbrace{\frac{\hat{\mathbf{p}}^2}{2m_e} + V(\mathbf{r})}_{\hat{H}_{\text{matter}}} + \underbrace{\sum_{\lambda} \hbar\omega_\lambda (\hat{a}^\dagger_\lambda \hat{a}_\lambda + \frac{1}{2})}_{\hat{H}_{\text{light}}} + \underbrace{\frac{q}{2m_e}(\hat{\mathbf{p}} \cdot \hat{\mathbf{A}} + \hat{\mathbf{A}} \cdot \hat{\mathbf{p}}) + \frac{q^2 \hat{\mathbf{A}} \cdot \hat{\mathbf{A}}}{2m_e}}_{\hat{W} = \hat{H}_{\text{light-matter}}}.} \qquad (29.80)$$

The Hamiltonian is now in a suggestive form, showing the "matter" part for the electron, the "light" part for the electromagnetic radiation

[26] The proton's effect on the electron energy is of course captured in the $V(\mathbf{r})$ term.

field, and the term $\hat{W}$ that couples the two. This equation is the starting point for the study of light-matter interaction. For certain cases, the entire Hamiltonian can be diagonalized to obtain hybrid eigenstates of light and matter. Examples are excitons and polaritons. Such **quasi-particles** have names ending in "...ons", just as electrons, neutrons, protons, and photons. They take a life of their own and can behave as genuine particles. So it is very important that you search whether the Hamiltonian can be solved exactly before applying perturbation theory. If you can do it, you are in luck, because you have discovered a new particle in nature! Typically the diagonalization requires transformation to a suitable basis. If we fail to diagonalize the Hamiltonian, we do perturbation theory.

Light-matter interaction as a perturbation: The method of applying perturbation theory is familiar by now: we identify the part of the Hamiltonian for which we know exact solutions, and treat the rest as the perturbation. In Equation 29.80, we realize that we know the *exact* solutions to $\hat{H}_{\text{matter}}$ and $\hat{H}_{\text{light}}$ independently. Let the eigenstate solutions for them be $\hat{H}_{\text{matter}}|\psi_e\rangle = E_e|\psi_e\rangle$ with E_e the electron eigenvalues, and $\hat{H}_{\text{light}}|\psi_{ph}\rangle = E_{ph}|\psi_{ph}\rangle$. Without the term $\hat{W}$ in Equation 29.80 there is no coupling between them because $\hat{\mathbf{p}}$ commutes with $\hat{a}_\lambda$ and $\hat{a}^\dagger_\lambda$. The annihilation and creation operators $\hat{a}_\lambda$ and $\hat{a}^\dagger_\lambda$ act *exclusively* on the photon Fock states, and do nothing to the $\hat{\mathbf{p}}$ and the $\mathbf{r}$ of electron states. That means $[\hat{H}_{\text{matter}}, \hat{H}_{\text{light}}] = 0$, i.e., they commute. Thus, the net quantum system can exist in a simultaneous eigenstate of both Hamitonians. It is then prudent to consider the net *unperturbed* Hamiltonian to be the sum $\hat{H}_0 = \hat{H}_{\text{matter}} + \hat{H}_{\text{light}}$, with the eigenstates the products $|\psi_e\rangle|\psi_{ph}\rangle = |\psi_e;\psi_{ph}\rangle = |e;n_1,n_2,...,n_\lambda,...\rangle$. Note that we have used the Fock-space or occupation number representation of the photon field in this notation. Then

$$\hat{H}_{\text{matter}}|e;n_1,n_2,...,n_\lambda,...\rangle = E_e|e;n_1,n_2,...,n_\lambda,...\rangle \text{ and}$$
$$\hat{H}_{\text{light}}|e;n_1,n_2,...,n_\lambda,...\rangle = \sum_\lambda \hbar\omega_\lambda(n_\lambda + \frac{1}{2})|e;n_1,n_2,...,n_\lambda,...\rangle \text{ give}$$
$$\hat{H}_0|e;n_1,n_2,...,n_\lambda,...\rangle = [E_e + \sum_\lambda \hbar\omega_\lambda(n_\lambda + \frac{1}{2})]|e;n_1,n_2,...,n_\lambda,...\rangle. \tag{29.81}$$

We focus our attention on the perturbation term $\hat{W}$ now. The vector potential field operator $\hat{\mathbf{A}}$ in this term in Equation 29.80 is a sum over all modes λ, just as in $\hat{H}_{\text{light}}$, and was derived explicitly in Equation 29.67. It is written as a field operator with photon creation and annihilation operators for each mode λ:

$$\boxed{\hat{\mathbf{A}}(\mathbf{r},t) = \sum_\lambda \hat{\mathbf{A}}_\lambda = \sum_\lambda \sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}\hat{\mathbf{e}}_\lambda[\hat{a}_\lambda e^{i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)} + \hat{a}^\dagger_\lambda e^{-i(\mathbf{k}_\lambda\cdot\mathbf{r}-\omega_\lambda t)}].} \tag{29.82}$$

The last term is $\frac{q^2\hat{\mathbf{A}}\cdot\hat{\mathbf{A}}}{2m_e} = \sum_\lambda[\frac{q^2\hbar}{2\epsilon_0\Omega m_e\omega_\lambda}(\hat{a}^\dagger_\lambda\hat{a}_\lambda + \frac{1}{2}) + \frac{q^2\hbar}{4\epsilon_0\Omega m_e\omega_\lambda}(\hat{a}^2_\lambda e^{2i\theta_\lambda} + (\hat{a}^\dagger_\lambda)^2 e^{-2i\theta_\lambda})]$, where $\theta_\lambda = \mathbf{k}_\lambda\cdot\mathbf{r} - \omega_\lambda t$. The first term of this expansion modifies the energy dispersion of the photon part of the unperturbed Hamiltonian. The second term causes the photon number to change by ±2, because of the $\hat{a}^2_\lambda$ and $(\hat{a}^\dagger_\lambda)^2$ terms (these are nonlinear processes). This entire interaction is non-linear, and is small for typical field strengths. So we will neglect this entire term in the following discussion, but the reader is encouraged to explore its consequences.

The rest of the perturbation for mode λ is $\hat{W}_\lambda = \frac{q}{2m_e}(\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}_\lambda + \hat{\mathbf{A}}_\lambda\cdot\hat{\mathbf{p}})$. Because of our choice of the Coulomb gauge $\nabla\cdot\hat{\mathbf{A}}_\lambda = 0$, for any scalar function f we have $\hat{\mathbf{p}}\cdot(\hat{\mathbf{A}}_\lambda f) = \hat{\mathbf{A}}_\lambda\cdot(\hat{\mathbf{p}}f) + f(\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}_\lambda)$. The last term is zero because $\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}_\lambda = -i\hbar(\nabla\cdot\hat{\mathbf{A}}_\lambda) = 0$. Thus, $\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}_\lambda = \hat{\mathbf{A}}_\lambda\cdot\hat{\mathbf{p}}$, and $[\hat{\mathbf{p}},\hat{\mathbf{A}}_\lambda] = 0$, i.e., $\hat{\mathbf{p}}$ and $\hat{\mathbf{A}}_\lambda$ commute, and the perturbation becomes $\hat{W}_\lambda = \frac{q}{2m_e}(\hat{\mathbf{p}}\cdot\hat{\mathbf{A}}_\lambda + \hat{\mathbf{A}}_\lambda\cdot\hat{\mathbf{p}}) = \frac{q}{m_e}\hat{\mathbf{A}}_\lambda\cdot\hat{\mathbf{p}}$. The perturbation then is written out explicitly in terms of the vector potential as[27]

$$\boxed{\hat{W}_\lambda = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}\Big(\hat{a}_\lambda e^{i\mathbf{k}_\lambda\cdot\mathbf{r}}\underbrace{e^{-i\omega_\lambda t}}_{\text{absorption}} + \hat{a}^\dagger_\lambda e^{-i\mathbf{k}_\lambda\cdot\mathbf{r}}\underbrace{e^{+i\omega_\lambda t}}_{\text{emission}}\Big)\hat{\mathbf{e}}_\lambda\cdot\hat{\mathbf{p}},} \tag{29.83}$$

where we identify the time-dependence of the oscillating perturbation. Now from our discussion of Fermi's golden rule for transition rates in Chapter 20, Section 20.2, Equation 20.23, we identify the $e^{-i\omega_\lambda t}$ term as responsible for the electron energy change $E_f = E_i + \hbar\omega_\lambda$, which is an absorption process. Similarly, the $e^{+i\omega_\lambda t}$ term is responsible for emission processes, for which $E_f = E_i - \hbar\omega_\lambda$. Fermi's golden rule for the transition rates is then captured in the equations

$$\boxed{\frac{1}{\tau_{abs}} \approx \frac{2\pi}{\hbar}\times|\langle f|\hat{W}^{abs}_\lambda|i\rangle|^2\delta[E_f - (E_i + \hbar\omega_\lambda)],} \tag{29.84}$$

and

$$\boxed{\frac{1}{\tau_{em}} \approx \frac{2\pi}{\hbar}\times|\langle f|\hat{W}^{em}_\lambda|i\rangle|^2\delta[E_f - (E_i - \hbar\omega_\lambda)].} \tag{29.85}$$

Let us first find the absorption matrix element. The perturbation term is

$$\hat{W}^{abs}_\lambda = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}e^{i\mathbf{k}_\lambda\cdot\mathbf{r}}\hat{a}_\lambda(\hat{\mathbf{e}}_\lambda\cdot\hat{\mathbf{p}}), \tag{29.86}$$

which we sandwich between the combined electron-photon *unperturbed* states $\langle e_f; m_1, m_2, ..., m_\lambda, ...|\hat{W}^{abs}_\lambda|e_i; n_1, n_2, ..., n_\lambda, ...\rangle$ to obtain the matrix element, which is

$$\langle f|\hat{W}^{abs}_\lambda|i\rangle = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}\langle e_f; m_1, m_2, ..., m_\lambda, ...|e^{i\mathbf{k}_\lambda\cdot\mathbf{r}}\hat{a}_\lambda(\hat{\mathbf{e}}_\lambda\cdot\hat{\mathbf{p}})|e_i; n_1, n_2, ..., n_\lambda, ...\rangle, \tag{29.87}$$

---

[27] Compared to the semi-classical approach of Chapter 27 where the perturbation was $\hat{W} = \frac{q}{m_e}\mathbf{A}\cdot\hat{\mathbf{p}}$ where the vector potential $\mathbf{A}$ was not an operator, now $\hat{W} = \frac{q}{m_e}\hat{\mathbf{A}}\cdot\hat{\mathbf{p}}$ has $\hat{\mathbf{A}}$ as an operator because of field quantization.

where it is clear that the light polarization vector $\hat{\mathbf{e}}_\lambda$ is a constant for a mode, the term $e^{i\mathbf{k}_\lambda\cdot\mathbf{r}}\hat{\mathbf{p}}$ acts *only* on the electronic part of the state, and the annihilation operator $\hat{a}_\lambda$ acts *only* on the photon part of the state. The matrix element thus factors into

$$\langle f|\hat{W}^{abs}_\lambda|i\rangle = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}\hat{\mathbf{e}}_\lambda \cdot \underbrace{\langle e_f|e^{i\mathbf{k}_\lambda\cdot\mathbf{r}}\hat{\mathbf{p}}|e_i\rangle}_{\mathbf{p}_{if}} \times \underbrace{\langle m_1, m_2, ..., m_\lambda, ...|\hat{a}_\lambda|n_1, n_2, ..., n_\lambda, ...\rangle}_{\sqrt{n_\lambda}\delta_{m_\lambda, n_\lambda - 1}}, \tag{29.88}$$

where we identify the momentum matrix element as $\mathbf{p}_{if}$ between the initial and final electronic states. This is a *vector* which includes the photon momentum through the $\mathbf{k}_\lambda$ term. The photon matrix element has the annihilation operator $\hat{a}_\lambda$ destroying one photon *only* in mode λ leaving all other modes unaltered, leading to $\langle ..., m_\lambda, ...|\hat{a}_\lambda|..., n_\lambda, ...\rangle = \sqrt{n_\lambda}\langle ..., m_\lambda, ...|..., n_\lambda - 1, ...\rangle = \sqrt{n_\lambda}\delta_{m_\lambda, n_\lambda - 1}$ because of the orthogonality of the photon modes. It is clear that if the electromagnetic field was in the vacuum state, the matrix element is zero because $\hat{a}_\lambda|0\rangle = 0$. This is just another way of saying that there cannot be an absorption event if there is nothing to absorb. In general (including when $n_\lambda = 0$) then, the matrix element is given by

$$\boxed{\langle f|\hat{W}^{abs}_\lambda|i\rangle = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}}(\hat{\mathbf{e}}_\lambda \cdot \mathbf{p}_{if})\sqrt{n_\lambda} \cdot \delta_{m_\lambda, n_\lambda - 1},} \tag{29.89}$$

and it enters Fermi's golden rule as the absolute square

$$|\langle f|\hat{W}_\lambda|i\rangle_{abs}|^2 = \frac{q^2}{m_e^2}\frac{\hbar}{2\epsilon_0\Omega\omega_\lambda}(\hat{\mathbf{e}}_\lambda \cdot \mathbf{p}_{if})^2 n_\lambda, \tag{29.90}$$

which is proportional to the photon number n_λ in mode λ. This proportionality is of utmost importance. For the momentum matrix element, we should retain the photon wavevector $\mathbf{k}_\lambda$ when necessary – for example in the Compton effect of scattering of a photon off an electron. However, for many cases of interest, the photon momentum is much smaller than that of the electron, and may be neglected. For those cases, $e^{i\mathbf{k}_\lambda\cdot\mathbf{r}} = 1 + i\mathbf{k}_\lambda \cdot \mathbf{r} + ... \approx 1$ is called the **dipole approximation**. With this approximation, we write the momentum matrix element in a modified form. We have $\langle e_f|\hat{\mathbf{p}}|e_i\rangle = m_e\langle e_f|\frac{d\hat{\mathbf{r}}}{dt}|e_i\rangle = -\frac{im_e}{\hbar}\langle e_f|\hat{\mathbf{r}}\hat{H}_0 - \hat{H}_0\hat{\mathbf{r}}|e_f\rangle = \frac{im_e}{\hbar}(E_{ef} - E_{if})\langle e_f|\hat{\mathbf{r}}|e_f\rangle = im_e\omega_{fi}\mathbf{r}_{if}$, where $\hbar\omega_{fi} = E_{ef} - E_{ei}$ is the change in the electron energy eigenvalue in the absorption event, and $\mathbf{r}_{if} = \int d^3r\psi^\star_{ef}(r)\mathbf{r}\psi^\star_{ei}(r)$ is the **dipole matrix element** which is obtained using the known electron eigenstates. Then, the square of the absorption matrix element is written compactly as

$$\boxed{|\langle f|\hat{W}^{abs}_\lambda|i\rangle|^2 = \frac{q^2 n_\lambda}{2\epsilon_0\Omega}\frac{\hbar\omega_{fi}^2}{\omega_\lambda}(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2.} \tag{29.91}$$

Following the same procedure, we compute the matrix element for *emission* of photons by electrons. The perturbation term is

$$\hat{W}_{\lambda}^{em} = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_{\lambda}}}e^{-i\mathbf{k}_{\lambda}\cdot\mathbf{r}}\hat{a}_{\lambda}^{\dagger}(\hat{\mathbf{e}}_{\lambda}\cdot\hat{\mathbf{p}}), \tag{29.92}$$

which has $\hat{a}_{\lambda}^{\dagger}$, the *creation* operator for photons. The emission matrix element is

$$\langle f|\hat{W}_{\lambda}^{em}|i\rangle = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_{\lambda}}}\hat{\mathbf{e}}_{\lambda}\cdot\underbrace{\langle e_f|e^{-i\mathbf{k}_{\lambda}\cdot\mathbf{r}}\hat{\mathbf{p}}|e_i\rangle}_{\mathbf{p}_{if}}\times$$
$$\underbrace{\langle m_1,m_2,...,m_{\lambda},...|\hat{a}_{\lambda}^{\dagger}|n_1,n_2,...,n_{\lambda},...\rangle}_{\sqrt{n_{\lambda}+1}\delta_{m_{\lambda},n_{\lambda}+1}}, \tag{29.93}$$

where the matrix element of the creation operator now gives something different from the annihilation operator: $\langle...,m_{\lambda},...|\hat{a}_{\lambda}^{\dagger}|...,n_{\lambda},...\rangle = \sqrt{n_{\lambda}+1}\langle...,m_{\lambda},...|...,n_{\lambda}+1,...\rangle = \sqrt{n_{\lambda}+1}\delta_{m_{\lambda},n_{\lambda}+1}$. This is from the stepping property of the creation operator. The matrix element for emission is then

$$\boxed{\langle f|\hat{W}_{\lambda}^{em}|i\rangle = \frac{q}{m_e}\sqrt{\frac{\hbar}{2\epsilon_0\Omega\omega_{\lambda}}}(\hat{\mathbf{e}}_{\lambda}\cdot\mathbf{p}_{if})\sqrt{n_{\lambda}+1}\cdot\delta_{m_{\lambda},n_{\lambda}+1}.} \tag{29.94}$$

With the dipole approximation, the square of the emission matrix element becomes

$$\boxed{|\langle f|\hat{W}_{\lambda}^{em}|i\rangle|^2 = \frac{q^2(n_{\lambda}+1)}{2\epsilon_0\Omega}\frac{\hbar\omega_{fi}^2}{\omega_{\lambda}}(\hat{\mathbf{e}}_{\lambda}\cdot\mathbf{r}_{if})^2.} \tag{29.95}$$

We note that the square of the emission matrix element[28] is proportional *not* to the photon number n_{λ} in mode λ, but to the photon number plus 1: $n_{\lambda}+1$. On this crucial difference between the emission and absorption processes rests the explanation of spontaneous emission. What it really means is that an emission process is allowed even when there are no photons ($n_{\lambda}=0$) in the electromagnetic field. Let us use it to evaluate the transition rates to see how this comes about.

The transition rate for photon absorption is obtained by using Equation 29.91 in Fermi's golden rule Equation 29.84 for absorption from a *single* photon mode λ:

$$\frac{1}{\tau_{i\to f}^{abs}(\omega_{\lambda})} = \frac{2\pi}{\hbar}|\langle f|\hat{W}_{\lambda}^{abs}|i\rangle|^2\delta[E_f-(E_i+\hbar\omega_{\lambda})]. \tag{29.96}$$

Since there are many modes available in the electromagnetic spectrum to cause this transition, the net rate should be the sum of the individual rates due to each mode λ:

$$\frac{1}{\tau_{i\to f}^{abs}} = \sum_{\lambda}\frac{1}{\tau_{i\to f}^{abs}(\omega_{\lambda})} = \sum_{\lambda}\frac{\pi q^2 n_{\lambda}}{\epsilon_0\Omega}\frac{\omega_{fi}^2}{\omega_{\lambda}}(\hat{\mathbf{e}}_{\lambda}\cdot\mathbf{r}_{if})^2\delta[E_f-(E_i+\hbar\omega_{\lambda})], \tag{29.97}$$

[28] Sakurai, in his book "Advanced Quantum Mechanics" has this to say about Equations 29.89, 29.91, 29.94 and 29.95 of absorption and emission matrix elements: "This simple rule is one of the most important results of this book".

where we have substituted Equation 29.91. The sum over the photon modes is recognized as an integration over the photon density of states computed by the recipe $\sum_\lambda(...) \to \int d(\hbar\omega_\lambda) D_E(\hbar\omega_\lambda)(...)$, $D_E(\hbar\omega_\lambda) = \frac{\Omega\omega_\lambda^2}{\hbar\pi^2c^3}$, which was obtained in Equation 29.46. This summation process thus yields the net absorption rate

$$\frac{1}{\tau_{i\to f}^{abs}} = \int d(\hbar\omega) D_E(\hbar\omega_\lambda) \frac{\pi q^2 n_\lambda}{\epsilon_0 \Omega} \frac{\omega_{fi}^2}{\omega_\lambda} (\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2 \delta[E_f - (E_i + \hbar\omega_\lambda)], \tag{29.98}$$

which simplifies to

$$\frac{1}{\tau_{i\to f}^{abs}} = \frac{q^2}{\pi\epsilon_0\hbar c^3} \int d\omega_\lambda \cdot (\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2 n_\lambda \omega_{fi}^2 \omega_\lambda \delta[(\omega_f - \omega_i) - \omega_\lambda)]. \tag{29.99}$$

Note that the macroscopic volume Ω has cancelled out, as it should. The Dirac delta function ensures energy conservation by picking out only those modes ω_λ of the photon field that satisfy $\omega_{fi} = \omega_f - \omega_i = \omega_\lambda$, where $E_{ef} = \hbar\omega_f$ and $E_{ei} = \hbar\omega_i$ are the final and initial *electron* energies in the transition. Note that the factor $(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2 = r_{if}^2 \cos^2\theta_\lambda$ where θ_λ is the angle between the dipole vector $\mathbf{r}_{if}$ and the mode polarization unit vector $\hat{\mathbf{e}}_\lambda$ changes as we sample the photon mode space. Similarly, the number of photons in mode λ also changes. Let us label the *average* contribution after the integration by $\overline{n_\lambda(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2} = n(\omega_{fi})\overline{(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2}$, which we will evaluate shortly. We then obtain the net absorption rate to be

$$\boxed{\frac{1}{\tau_{i\to f}^{abs}} = \frac{e^2\omega_{fi}^3}{\pi\epsilon_0\hbar c^3} \overline{(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2} n(\omega_{fi}) = \frac{1}{\tau_0} n(\omega_{fi}).} \tag{29.100}$$

Here we have defined a characteristic transition rate $\frac{1}{\tau_0} = \frac{e^2\omega_{fi}^3}{\pi\epsilon_0\hbar c^3}\overline{(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2}$. By following exactly the same procedure, we obtain the net emission rate to be

$$\boxed{\frac{1}{\tau_{i\to f}^{em}} = \frac{1}{\tau_0}[n(\omega_{fi}) + 1] = \frac{1}{\tau_0}[\underbrace{n(\omega_{fi})}_{\text{stimulated}} + \underbrace{1}_{\text{spontaneous}}].} \tag{29.101}$$

29.17 Fundamental optical processes

Equations 29.100 and 29.101 capture much of the important physics of photon absorption and emission by matter, and consequently that of light-matter interaction. The rate of photon absorption $\frac{1}{\tau_{i\to f}^{abs}} = \frac{1}{\tau_0} n(\omega_{fi})$ is proportional to the number of photons already present in the photon field, which is reasonable. This process was discussed in detail in Chapter 27, and is responsible for semiconductors photodetectors and solar cells discussed in Chapter 28.

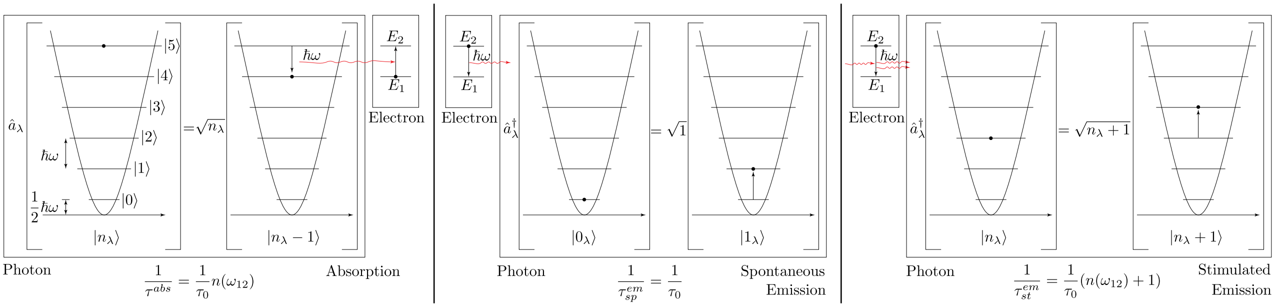

$|5\rangle$
$|4\rangle$
$|3\rangle$
$|2\rangle$
$|1\rangle$
$|0\rangle$
$\hat{a}_\lambda$
$\hbar\omega$
$\frac{1}{2}\hbar\omega$
$|n_\lambda\rangle$
$=\sqrt{n_\lambda}$
$|n_\lambda - 1\rangle$
$\hbar\omega$
E_2
E_1
Electron
Photon
$\frac{1}{\tau^{abs}} = \frac{1}{\tau_0} n(\omega_{12})$
Absorption
E_2
$\hbar\omega$
E_1
Electron
$\hat{a}^\dagger_\lambda$
$= \sqrt{1}$
$|0_\lambda\rangle$
$|1_\lambda\rangle$
Photon
$\frac{1}{\tau^{em}_{sp}} = \frac{1}{\tau_0}$
Spontaneous Emission
E_2
$\hbar\omega$
E_1
Electron
$\hat{a}^\dagger_\lambda$
$= \sqrt{n_\lambda + 1}$
$|n_\lambda\rangle$
$|n_\lambda + 1\rangle$
Photon
$\frac{1}{\tau^{em}_{st}} = \frac{1}{\tau_0}(n(\omega_{12}) + 1)$
Stimulated Emission

Fig. 29.37 Illustration of the processes of absorption, spontaneous emission, and stimulated emission from the perspective of a quantized electromagnetic field.

The earlier part of this chapter discussed emission processes $\frac{1}{\tau^{em}_{i\to f}} = \frac{1}{\tau_0}[n(\omega_{fi}) + 1]$, which is is proportional to $n(\omega_{fi}) + 1$. This implies that an electron in an excited state in an atom is perfectly capable of emitting a photon into vacuum – and populate the vacuum state $|0, 0, ...\rangle$ with $n(\omega_{fi}) = 0$ with a photon and cause it to transition to the state $|0, ..., 0, 1_\lambda, 0, ...\rangle$. What this means is that an electron in the excited state is capable of emitting a photon even when there is no electromagnetic radiation incident on it. An explanation of this process of *spontaneous* emission is fundamentally impossible from classical electrodynamics because when there is no radiation field, $\mathbf{A} = 0$, and there is no light-matter interaction. It is also impossible to obtain from quantum mechanics if one quantizes just the electron part of the Hamiltonian, and not the electromagnetic field – in a semi-classical picture[29]. That is because in the absence of electromagnetic radiation, again $\mathbf{A} \cdot \hat{\mathbf{p}} = 0$, and an electron in an eigenstate is a stationary state – it simply has no reason to change its energy. The simple explanation of spontaneous emission obtained by quantization of the electromagnetic field $\mathbf{A} \to \hat{\mathbf{A}}$ is a triumph of quantum field theory. The process of spontaneous emission is responsible for photon emission from light bulbs and semiconductor light-emitting diodes (LEDs).

Since the emission rate is $\frac{1}{\tau^{em}_{i\to f}} = \frac{1}{\tau_0}[n(\omega_{fi}) + 1]$, the emission rate is *increased* if there are already photons $n(\omega_{fi})$ in the modes into which the electron is allowed to emit. Thus, the presence of photons in the field *stimulates* the electron to emit more, which is the reason for calling the process *stimulated* emission. This process produces optical gain, and was central to the realization of the laser discussed earlier in this chapter. Thus the laser, the precious child of field quantization, and the driver of quantum electronics, has revolutionized many areas of science, engineering, and applications.

Fig. 29.37 summarizes our results for the fundamental processes of light-matter interaction. Note that when the (electron+photon) system is considered as one bigger system with internal energies, the processes of absorption, spontaneous emission, and stimulated emission

[29] If the electron part of the Hamiltonian is *also* a harmonic oscillator, however, it is indeed possible to obtain a spontaneous emission lifetime because of a crucial cancellation of $\hbar$! The physical meaning of the lifetime is the radiation damping of a charged harmonic oscillator.

are seen to be natural and necessary to conserve energy. Let us discuss the absorption and emission rates quantitatively, and get a feel for the timescales of optical transitions. The characteristic timescale is $\frac{1}{\tau_0} = \frac{q^2\omega_{fi}^3}{\pi\epsilon_0\hbar c^3}\overline{(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2}$. To evaluate the angular factor $\overline{(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2}$, let us orient the dipole vector $\mathbf{r}_{if} = r_{if}\hat{\mathbf{z}}$ along the z-axis, and consider a situation when the allowed mode polarization vectors $\hat{\mathbf{e}}_\lambda$ vary isotropically. Since $\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if} = r_{if}\cos\theta$, we obtain the average over solid angles[30] Ω in spherical coordinates $d\Omega = \sin\theta d\theta d\phi$ to be

$$\overline{(\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2} = \frac{\int d\Omega (\hat{\mathbf{e}}_\lambda \cdot \mathbf{r}_{if})^2}{\int d\Omega} = r_{if}^2 \frac{\int_{\phi=0}^{2\pi} d\phi \int_{\theta=0}^{\pi} d\theta \sin\theta (\cos\theta)^2}{\int_{\phi=0}^{2\pi} d\phi \int_{\theta=0}^{\pi} d\theta \sin\theta} = \frac{1}{3} r_{if}^2, \tag{29.102}$$

and the characteristic rate thus becomes

$$\boxed{\frac{1}{\tau_0} = \frac{q^2\omega_{fi}^3 r_{if}^2}{3\pi\epsilon_0\hbar c^3} \approx (\frac{900 \text{ nm}}{\lambda})^3 \times (\frac{r_{if}}{1 \text{ nm}})^2 \times (10^9 \text{ sec}),} \tag{29.103}$$

where we have used $\omega_{fi} = \frac{2\pi c}{\lambda}$, and scaled the wavelength and the dipole matrix element to characteristic scales. For a wavelength of 900 nm and a dipole length of 1 nm, $\tau_0 \approx 1$ ns. This is the spontaneous emission lifetime, because when $n(\omega_{fi}) = 0$, $\tau_{i\to f}^{em} = \tau_0$. The timescale we have found is very close to what is measured for the hydrogen atom: it takes an excited electron roughly a nanosecond to transition to the ground state and emit a photon spontaneously. However, note that the accurate value is sensitive to the frequency (or wavelength) of the photon, and the dipole matrix element. The nature of the 3D photon DOS makes the rate increase as the energy of transition increases. The stronger the dipole matrix element, the faster the transition.

29.18 Einstein's re-derivation of Planck's law: back to 1^2!

As a closure to the contents of this book, we return to where we started. Let us take a second look at how Einstein used the above picture of light-matter interaction to re-derive Planck's law of blackbody radiation. Our discussions of Planck's distribution in earlier chapters had lacked clarity of the nature of the "matter" and its temperature. We had alluded that field-quantization will resolve this confusion. We are ready now to see how.

Consider now that there are various kinds of atoms present in the cubic box we had used to index the modes of the electromagnetic field, as re-sketched in Fig. 29.38. The atoms are at temperature T. Electrons in the atoms can make transitions between various energies by either absorbing photons from, or emitting photons into the electromagnetic field, which co-exists, and is in *equilibrium* with the atoms. Of the various kinds of atoms, let there be N atoms of a particular species

[30] Here Ω is the solid angle, not the volume of the radiation field.

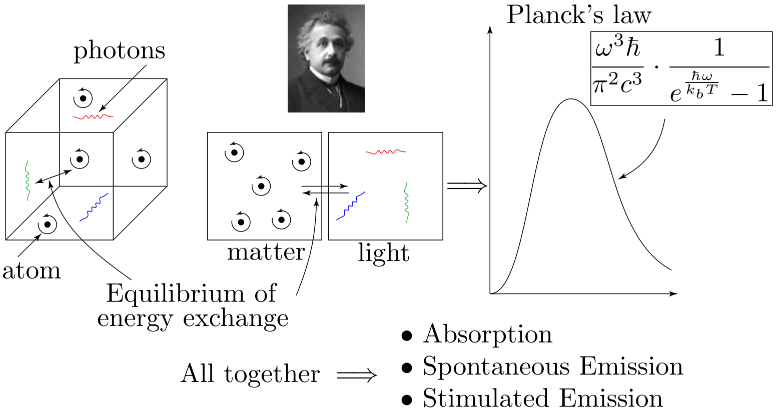


Fig. 29.38 Einstein's derivation of Planck's blackbody radiation law treated interaction of light and matter, and invoked the concepts of absorption, and spontaneous and stimulated emission.

for which the electron eigenvalues are E_2 and E_1, with $E_2 - E_1 = \hbar\omega$. From Boltzmann statistics for electrons, the N atoms are distributed into $N = N_1 + N_2$, where $\frac{N_2}{N_1} = e^{-\frac{E_2-E_1}{k_bT}} = e^{-\frac{\hbar\omega}{k_bT}}$. So $N_1 = \frac{e^{\frac{\hbar\omega}{k_bT}}}{e^{\frac{\hbar\omega}{k_bT}}+1}N$, and $N_2 = \frac{1}{e^{\frac{\hbar\omega}{k_bT}}+1}N$. The fact that the photons are in equilibrium with the atoms means that there is no net exchange of energy between them, which is possible if

$$N_1 \times (\text{Rate of energy absorption}) = N_2 \times (\text{Rate of energy emission}). \tag{29.104}$$

Now the net absorption rate of photons is given by Equation 29.100, i.e., $\frac{1}{\tau^{abs}} = \frac{n(\omega)}{\tau_0}$, and each photon absorption brings in an energy $\hbar\omega$. With the spontaneous and stimulated emission rates from Equation 29.101 we get

$$\left(\frac{e^{\frac{\hbar\omega}{k_bT}}}{e^{\frac{\hbar\omega}{k_bT}}+1}N\right) \times \left(\frac{n(\omega)\hbar\omega}{\tau_0}\right) = \left(\frac{1}{e^{\frac{\hbar\omega}{k_bT}}+1}N\right) \times \left(\frac{[n(\omega)+1]\hbar\omega}{\tau_0}\right)$$
$$\implies \boxed{n(\omega) = \frac{1}{e^{\frac{\hbar\omega}{k_bT}}-1},} \tag{29.105}$$

where the photon occupation number $n(\omega)$ in mode ω is obtained as the now-familiar Bose–Einstein distribution function. Now the values of E_2 and E_1 are not specific – meaning there are other levels available for single atoms, and there are many kinds of atoms. Thus, the electron energy differences for atoms span a range of $\hbar\omega$ values. In that case, the energy content in the photon field in equilibrium with the atoms is

$$u_\omega(\omega)d\omega = \frac{D_E(\hbar\omega)}{\Omega}d(\hbar\omega) \times [n(\omega)\hbar\omega] = \frac{\Omega\omega^2}{\Omega\hbar\pi^2c^3}(\hbar d\omega)\left(\frac{\hbar\omega}{e^{\frac{\hbar\omega}{k_bT}}-1}\right), \tag{29.106}$$

which immediately yields Planck's blackbody spectrum

$$\boxed{u_\omega(\omega) = \frac{\omega^3\hbar}{\pi^2 c^3}\frac{1}{e^{\frac{\hbar\omega}{kT}}-1}.} \tag{29.107}$$

This is the same result as before, but now the interaction of light and matter is made explicit by Einstein's derivation. Note that though we call this Einstein's derivation, he actually introduced phenomenological A and B coefficients in his 1917 paper and invoked the concepts of spontaneous and stimulated emissions based on pure reasoning. We discussed this in Section 29.2. It took at least 10 more years after his paper for the basics of quantum mechanics and perturbation theory to be in place, and then for Dirac to quantize the electromagnetic field and explain the reason for spontaneous emission. So the quantization of the electromagnetic field gave Einstein's heuristic derivation a rigorous quantitative underpinning, and enabled us to treat the subject with the benefit of hindsight.

29.19 Chapter summary section

In this chapter, we learned:

- Einstein's uncovering of the processes of absorption, spontaneous emission, and stimulated emission as the three fundamental optical processes,
- How population inversion is achieved in semiconductors by the injection of excess electrons and hole and splitting of the quasi-Fermi levels,
- How optical gain and optical cavities are combined with double heterostructures to make semiconductor lasers,
- Why Fabry–Perot cavity lasers are typically multimode, and how one can make single-mode lasers,
- The construction of LEDs and quantum cascade lasers, and
- How quantization of the electromagnetic field allows the direct calculation of the spontaneous emission rate, so we can understand the nature of photons.

Since all matter was produced from light in the Big Bang, immense energy is stored in every tiny bit of matter. The energy is stored in a very ordered fashion as the electronic energy in the separation of electron energy levels, and the energy in chemical bonds, the nuclear energy in the protons and neutrons in the nucleus, and the rest-mass energy as $E = mc^2$ of the electrons, protons, and the neutrons. Most photonic devices aim to use the transitions of electrons between allowed energy levels to absorb, detect, or to produce light of specific photon energies and phases. Transitions between electrons in the valence and conduction bands energy states of semiconductors can absorb or produce energies in the $\sim 0 - 10$ eV range. But transitions between the non-bonding electronic core energy levels can absorb and produce much higher and sharper energies in the X-ray region of the electromagnetic spectrum. These "high-energy" photonic phenomena are routinely used to measure chemical and structural properties of semiconductors, their surfaces, and defects in them.

With the discussion of this chapter, it seems we have come "full circle", back to where things started: Planck's blackbody radiation formula! The hope is that the journey has made us wiser and in the journey we have explored several classical and quantum properties of semiconductors and their nanostructures. We have seen what principles power their incredible utility in so many electronic and photonic device applications, a summary of which is given in the final chapter (Chapter 30). At the same time, semiconductors remain a hotbed for the exploration of fundamental physics and materials science itself. I am sure you are ready for the next trip to a higher mountain!

Further reading

The earliest books on semiconductor lasers were **Heterostructure Lasers** by H. C. Casey and M. B. Panish and **Semiconductor Lasers and Heterojunction LEDs** by H. Kressel and J. K. Butler, which remain highly relevant today because of their emphasis on the fundamentals. **Quantum-Confined Laser Devices: Optical gain and recombinations in semiconductors** by P. Blood has a modern and excellent pedagogical treatments of the subject of semiconductor lasers and optical properties. **Light Emitting Diodes** by F. Schubert has a modern treatment of the physics and technology of semiconductor LEDs. **Physics of Photonic Devices** by S. L. Chuang has a comprehensive treatment of the details of several aspects of the physics and device design and modeling of semiconductor lasers. **Physics of Semiconductor Laser Devices** by G. H. B. Thompson has a good treatment of the design and analysis of semiconductor laser devices, with interesting energy band diagram representations. **Semiconductor Lasers** by G. P. Agarwal and N. K. Dutta is an early and influential text on heterostructure lasers that are used in communication systems. **Diode Lasers and Photonic Integrated Circuits** by L. A. Coldren, S. W. Corzine, and M. L. Mashanovitch is a modern treatment with an excellent balance that exploits the combination of phenomenological models with deep underlying physics, and also gives typical state of the art semiconductor laser design rules, materials, and applications. **Laser Electronics** by J. Verdeyen is an excellent and readable book with well-written and illustrated chapters on semiconductor lasers, and their comparison to all other types of lasers. To learn the rich history of lasers and their physics seamlessly, **Masers and Lasers: An Historical Approach** by M. Bertolloti is highly recommended reading for both the laser enthusiast and the expert (may you be both!).

Exercises

(29.1) **Roosbroeck–Shockley relations**

(a) Prove the approximate form of Equation 29.19 for semiconductors.

(b) Plot the spectra of a semiconductor of bandgap $E_g = 1.4$ eV for various voltages.

(c) Comment on why typical spontaneous emission spectra can differ from the calculated values that assume perfect band-edge emission.

(d) Discuss the importance of this process to light-emitting diodes.

(29.2) **Population inversion, optical gain, and lasing**

In this chapter we derived that the equilibrium optical absorption coefficient of a semiconductor is $\alpha_0(\hbar\omega) = C_0|\hat{e} \cdot \mathbf{p_{cv}}|^2 \rho_r(\hbar\omega - E_g)$, where $\rho_r(\hbar\omega - E_g)$ is the joint electron-photon DOS, $C_0 = \frac{\pi e^2}{n_r c \epsilon_0 m_0^2 \omega}$, and all symbols have their usual meanings. We also discussed that under *non-equilibrium* conditions, the optical absorption coefficient becomes $\alpha(\hbar\omega) = \alpha_0(\hbar\omega)[f_v(k) - f_c(k)]$, where the electron occupation functions of the bands are given by the Fermi–Dirac distributions $f_v(k) = \frac{1}{1+e^{\frac{E_v(k)-F_v}{k_b T}}}$ and $f_c(k) = \frac{1}{1+e^{\frac{E_c(k)-F_c}{k_b T}}}$, but with quasi-Fermi levels F_v for the valence band and F_c for the conduction band as the mathematical means to capture *non-equilibrium* conditions. Consider a semiconductor with parabolic bandstructures for the conduction band $E_c(k) = E_g + \frac{\hbar^2 k^2}{2m_c^\star}$ and valence band $E_v(k) = -\frac{\hbar^2 k^2}{2m_v^\star}$ for this problem.

(a) Make a sketch of the equilibrium absorption coefficients $\alpha_0(\hbar\omega)$ for a bulk 3D semiconductor, and a 2D quantum well vs. the photon energy $\hbar\omega$.

(b) Plot the Fermi difference function $f_v(k) - f_c(k)$ as a function of the photon energy $\hbar\omega$ for a few choices of the quasi-Fermi levels F_v, F_c of the valence and conduction bands. Specifically, track the photon energy at which the difference function *changes sign* from negative to positive.

Fig. 29.39 Zores Alferov was awarded the 2000 Nobel Prize in Physics for the demonstration of the double heterojunction semiconductor laser.

(c) Now combine (a) and (b) to plot the non-equilibrium absorption coefficient $\alpha(\hbar\omega)$ for the choices of F_v, F_c from part (b). Discuss the significance of a *negative* absorption coefficient.

(d) Show that the requirement for population inversion is $F_c - F_v > E_c - E_v = \hbar\omega$. This is the famous Bernard–Duraffourg condition for *population inversion* in semiconductor lasers.

(e) Based on your results above, explain why semiconductor heterostructure quantum wells have lower injection thresholds for lasing than 3D bulk semiconductors. This idea and demonstration won Herbert Kroemer (Fig. 15.21) and Zhores Alferov (Fig 29.39) the 2000 Nobel Prize in Physics.

(29.3) **Gain in quantum wells**

Instead of bulk semiconductors, if we have a quantum well of thickness L_z, the absorption coefficient still follows the joint density of states feature. To derive this, do the following: assume an infinite barrier height (this approximation can be relaxed in a numerical approach, we keep things analytical here). Then,

(a) Write down the wavefunctions of electron states in the n_c^{th} subband if the conduction band $|n_c, \mathbf{k_c}\rangle$ as a product of the quantized envelope wavefunction $C(n_c, z)$ along the confinement (z-)direction, a free–electron envelope wavefunction $t_c(x, y, \mathbf{k_c})$ in the transverse $(x - y-)$direction, and a CB Bloch periodic function $u_c(\mathbf{r})$. Do the same for the valence band states $|n_v, \mathbf{k_v}\rangle$, choosing the corresponding envelope functions $V(n_v, z)$, $t_v(x, y, \mathbf{k_v})$, and Bloch periodic function $u_v(\mathbf{r})$.

(b) Argue why the transverse parts of the enveloped wavefunctions $t_c(x, y, \mathbf{k_c})$ and $t_v(x, y, \mathbf{k_v})$ must be the *same* for optical transitions.

(c) The optical matrix element is given by $W_{op} = \langle n_c, \mathbf{k_c}| - \frac{e}{m_0}\mathbf{A}\cdot\mathbf{p}|n_v, \mathbf{k_v}\rangle$. Show that this can be written as the product $W_{op} = \frac{eA_0}{m_0} \times \hat{e}\cdot\underbrace{\langle u_c|\mathbf{p}|u_v\rangle}_{\mathbf{p_{cv}}} \times \delta_{\mathbf{k_c},\mathbf{k_v}} \times I^{c,n_c}_{v,n_v}$, where the optical transition rules are dictated by the integral $I^{c,n_c}_{v,n_v} = \int_{-\infty}^{+\infty} C^\star(n_c, z)V(n_v, z)dz$.

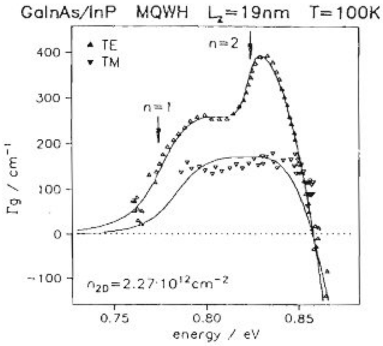


Fig. 29.40 Experimentally measured optical gain in an InP/InGaAs/InP quantum well heterostructure. From IEEE Journal of Quantum Electronics, vol. 25, 1407 (1989).

(d) Show that for infinite barrier quantum wells, the selection rule is $I^{c,n_c}_{v,n_v} = \delta_{n_v,n_c}$, i.e., the only transitions from valence band quantized states to conduction band quantized states that are allowed follow the selection rule $\Delta n = n_v - n_c = 0$ (for example, $1 \to 1, 2 \to 2$ ...).

(e) Based on part (d), show that the *equilibrium* absorption coefficient is then given by

$$\alpha_0(\hbar\omega) = \underbrace{\frac{\pi e^2}{n_r c\epsilon_0 m_0^2 \omega}|\hat{e}\cdot\mathbf{p_{cv}}|^2}_{same-coeff-as-3D} \times \underbrace{\frac{m_r^\star}{\pi\hbar^2 L_z}}_{\rho_r^{2D}} \times \underbrace{\sum_n \theta[\hbar\omega - (E_g + E_n^c + E_n^v)]}_{staircase-function}, \tag{29.108}$$

where E_n^c, E_n^v are the conduction and valence band quantization energies and $n = n_c = n_v$.

(f) Fig. 29.40 shows an experimentally measured optical gain spectrum of a lattice-matched InP/InGaAs/InP quantum well. Use the results of the this problem, and also Fig. 28.22 to explain the gain spectrum.

(29.4) **Threshold current of diode lasers**

Show that by writing the gain as $g = g_0 \ln(N/N_{tr})$, the threshold current of a diode laser with a electron cavity volume V_e, and in which the recombination is dominated by spontaneous recombination at the rate $R_{sp} = BN^2$ is given by

$$I_{th} = \frac{qV_e BN_{tr}^2}{\eta_i} e^{\frac{2(\alpha_i + \alpha_m)}{\Gamma g_0}}. \tag{29.109}$$

(29.5) **Blue laser diode**

Consider the edge-emitting Fabry–Perot cavity blue laser diode in Section 29.10.

(a) Find the differential efficiency η_d from Fig. 29.24 (b).

(b) Given that $\eta_i = 0.95$ and the measured loss is $\alpha_i = 18$ cm^{-1}, find the mirror loss α_m.

(c) From the mirror loss of part (b), find the cavity length L assuming cleaved facets are used for the device.

(d) Fig. 29.25 (d) shows the output spectra above threshold. Find the cavity length from this plot and verify that your calculation of part (d) is correct.

(e) From Fig. 29.26 identify the transparency current density J_{tr}.

(29.6) **Schawlow–Townes linewidth**

(a) Show that because the photon intensity in a laser cavity where pumping is turned off at $t = 0$ decays from its original value at time $t = 0$ as $I(t) = I(0)e^{-t/\tau_p}\theta(t)$ where $\theta(t)$ is the unit step function, the electric field decays as $E(t) = E_0 e^{i\omega_0 t}e^{-t/2\tau_p}\theta(t)$.

(b) Show that the square of the Fourier transform of the electric field is the Lorentzian function $|E(\omega)|^2 = \frac{|E(\omega_0)|^2}{1+(\omega-\omega_0)^2\cdot(2\tau_p)^2}$.

(c) By making a plot of this response, show that the full-width half-maximum linewidth is $\Delta\omega = \frac{1}{\tau_p}$.

(29.7) **Temperature variation of semiconductor lasers** The temperature dependence of the threshold current of semiconductor lasers is experimentally found to vary as $I_{th} = I_0 e^{T/T_0}$, where T_0 is a characteristic temperature. This causes some headache for users and therefore designers should make lasers that are insensitive to temperature.

(a) Explain the temperature dependence of the threshold current by considering the dependence of the semiconductor gain coefficient on temperature.

(b) Obtain an approximation relation for T_0.

(c) Suggest ways to make the T_0 high so that the thermal variation is lowered.

(29.8) **Quantum theory of an LC oscillator** (a) Follow Dirac's procedure and quantize the LC oscillator energy $H_{cl} = \frac{1}{2}LI^2 + \frac{1}{2}CV^2$. Identify the conjugate variables (q, p), and check if they satisfy Hamilton's equations $\frac{dq}{dt} = \frac{\partial H_{cl}}{\partial p}$ and $\frac{dp}{dt} = -\frac{\partial H_{cl}}{\partial q}$. Then quantize it by postulating $[\hat{q}, \hat{p}] = i\hbar$.

(b) The uncertainty in an observable p is defined *precisely* as $\Delta q = \sqrt{\langle q^2 \rangle - \langle q \rangle^2}$. Express the uncertainties in voltage ΔV and current ΔI, and the product $\Delta V \Delta I$ of the ideal LC circuit as a function of frequency ω and the impedance $Z = \sqrt{L/C}$, and provide numerical values of the uncertainties.

(29.9) **Spontaneous emission and its classical interpretation**

In this chapter we found an explanation for the reason for spontaneous emission of photons by electrons after quantizing the electromagnetic field. We found the rate for spontaneous emission to be $\frac{1}{\tau_0} = \frac{e^2 \omega_{fi}^3 r_{if}^2}{3\pi\epsilon_0 \hbar c^3}$, where the symbols have their usual meanings.

(a) Explain the reasoning and describe the difference between absorption, spontaneous emission, and stimulated emission.

(b) Show that if the spontaneous emission lifetime for an electron transitioning from the first excited state to the ground state of a 1D harmonic oscillator is $\tau_{HO} = \frac{6\pi\epsilon_0 m_e c^3}{e^2 \omega_{fi}^2}$. Note that $\hbar$ does not appear in this result. This means one can potentially obtain this result from classical electrodynamics. Refer to *Classical Electrodynamics* by J. D. Jackson, Chapter 16 for an explanation, based on the radiation damping of a charged harmonic oscillator.

(c) Show that the field operators for the vector potential $\hat{\mathbf{A}}$ and the electric $\hat{\mathbf{E}}$ fail to commute, i.e., $[\hat{\mathbf{A}}, \hat{\mathbf{E}}] \neq 0$. Calculate $\langle \hat{\mathbf{E}} \rangle$ and $\langle \hat{\mathbf{E}}^2 \rangle$, and estimate the value of $\sqrt{\langle \hat{\mathbf{E}}^2 \rangle}$, the zero-point fluctuation of the electric field in vacuum.

(29.10) **Tunnel-injection into a quantum dot laser**

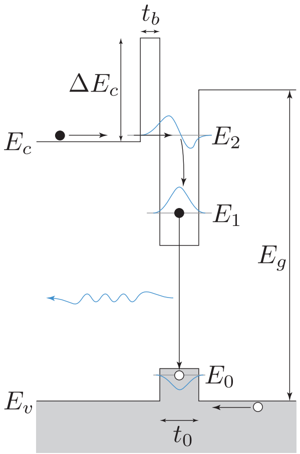


Fig. 29.41 Energy band diagram illustrating tunneling injection into a semiconductor heterostructure quantum dot laser.

In this problem, we consider a problem of tunneling injection of electrons into a quantum dot laser, which captures several concepts developed in this book, and thus is a good example to end the book. Fig. 29.41 shows the energy band diagram of a quantum dot of size t_0 in all three dimensions (typically $t_0 \sim$ nms). It is surrounded by wider bandgap semiconductors. We want to inject electrons into the conduction states of the dot, and holes into the valence states of the dot. We want the electron and hole then to recombine to emit a photon of a specific color. This region will then be placed within mirrors to make a laser, but for this problem, let us track the physics of semiconductor nanostructures till the emission of the photon.
(a) Based on your knowledge in this course, say a few words on how such a semiconductor heterostructure is actually realized: what sorts of semiconductor materials and doping are neces-

sary, and how you will create the desired structure for any photon wavelength of your choice.

(b) Assume all conduction effective masses to be $m_c^\star$ and valence effective masses be $m_v^\star$. Assuming the dot is of size t_0 in all three dimensions and the barrier heights in the conduction and valence bands are infinite, find the energies of the first and second quantized states shown in the conduction band, and the one quantized ground state shown in the valence band.

(c) If the barrier height is finite, is the second conduction state in the dot a true bound state? Why or why not? What about the ground state?

(d) Driven by available band alignments and requirement of high speed, I decide to inject electrons to the first excited state instead of the ground state. Give an expression for the tunneling probability of electrons from the conduction band on the left into the dot in terms of the parameters in the figure.

(e) After the electron is in the second conduction state, it has two ways to lose energy: heat, or light. Discuss why if the energy difference $E_2 - E_1$ is not equal to the energy of an optical phonon $\hbar\omega_o$, it is difficult for the electron to relax from E_2 to E_1 because it is in a quantum dot (0D). Explain why such a relaxation would not be a problem in a quantum well (2D) or a quantum wire (1D).

(f) Now using Fermi's golden rule explain why it is difficult for the electron in state E_2 to transition to the hole state E_0 to emit light. Remember the potential due to light or photons seen by electrons is of the type $W(x) \approx eFx$ in the dipole approximation.

(g) This deadlock is called the *phonon bottleneck*. When such heterostructures are realized, what is observed is that the electron does make it to state E_1 by scattering off with the hole in the valence band. Explain why this can resolve the deadlock and make it possible for the photon to be emitted.

Every End is a New Beginning

30

In this last chapter, we first step back to take a broad view of the current technological areas of application of semiconductors that are rooted in the physics of electrons and holes that we have discussed in the book. Then, we discuss what the future may hold in this field. This chapter is thus organized in the following way:

- We first provide a broad summary of the applications of semiconductors today, current challenges, and the state of the art.
- We then place them in a broader scientific context to look back and then look forward to possibilities in the future.
- We then discuss several new avatars of semiconductors, where they either exhibit new functionalities by themselves, or by close integration with other materials.
- We end this book with a discussion of the role of semiconductors in sectors of energy, healthcare, space exploration, and in societal changes in the past, present, and the future.

30.1 Smallest, fastest, brightest 829
30.2 Looking back and forward 833
30.3 Ferroelectric semiconductors 834
30.4 Ferromagnetic semiconductors 835
30.5 Multiferroic semiconductors 836
30.6 Superconducting semiconductors 836
30.7 Semiconductors for quantum communications 837
30.8 Semiconductors for quantum computation 840
30.9 Semiconductors for energy 842
30.10 Semiconductors for healthcare and agriculture 843
30.11 Semiconductors for space exploration 844
30.12 Social impact of semiconductors 844
30.13 Chapter summary section 845
Further reading 846
Exercises 846

30.1 Smallest, fastest, brightest

Today semiconductors are synonymous with **nanotechnology**. Moore's law has driven down the size of silicon transistors (**Chapter 19**) to ~ 5 nm in size, and $\sim$ 10s of billions of them are packed into a square centimeter chip. These microprocessors today are the electronic brains that handle a large number of complex computational tasks ranging from financial transactions and stock markets, security of nations and government functions, medical records and diagnostics, transportation networks, social networks, electric and power grids, and so much more. This drive of CMOS technology has created highly refined tools for fabrication, metrology, testing, and packaging of complex microprocessors, and a global supply chain. The production of silicon chips is an international effort: it is said that today a grain of sand (which is the source of silicon) makes two trips around the globe through several chemical and physical processing steps before it becomes a part of a silicon microprocessor. The fabrication and metrology tools and processes developed for microprocessors are being heavily leveraged to build the next generation of quantum computers (Fig. 30.1).

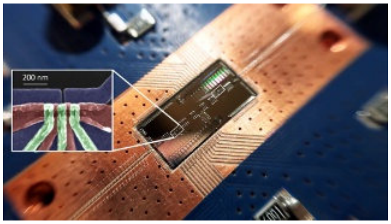


Fig. 30.1 Silicon spin qubits using lithographically defined nanoscale quantum dots. Shared by Dr. Felix Borjans and Dr. J. Petta, Princeton University.

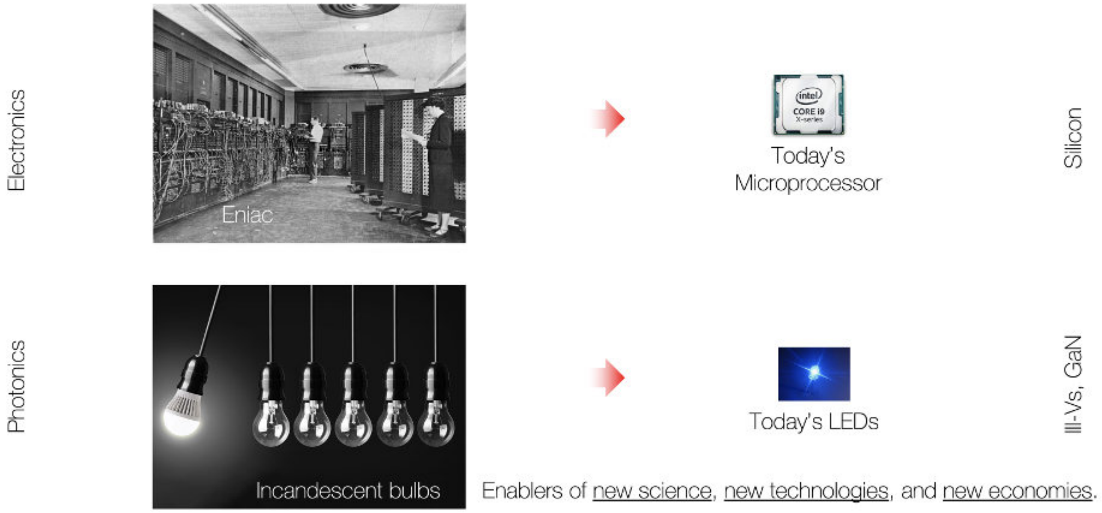


Fig. 30.2 The impact of semiconductors on the reduction of size, weight, power, and cost of electronic computation and solid-state lighting. In addition to fundamentally changing electronic and photonics, they are also the drivers of new and unanticipated functionalities.

Memory technology is undergoing a rapid expansion to store the vast amounts of data that is being created. In addition to traditional magnetic memory, an enormous expansion is being driven by silicon based flash technology (**Chapter** 24). This technology enabled the mobile and handheld device revolution in the last decade by offering a lightweight and energy efficient alternative to magnetic hard drives. Denser and more intimate embedding of memory within the logic and computation units of the microprocessor are now desired to enable efficient artificial intelligence algorithms and to give processors more (machine-) learning capacity. These approaches are also investigating the integration of non-semiconductor memories such as magnetic tunnel junctions with silicon CMOS.

While silicon and its uses are well known and publicized, it is not as well known that HEMT amplifiers made of GaAs, InP, or GaN semiconductors in mobile phones, cell-phone base stations, and power converters today also use 2D electron gases in III-V quantum wells that are controlled at the nanoscale (**Chapters** 19, 23). Some of these transistors today reach speeds exceeding 1 THz at room temperature and are enabling ultra high-frequency communication systems. Fast, high-voltage transistors made of GaN HEMTs are enabling significant shrinking of high-voltage power convertors for energy-efficient power electronics.

Transistors and diodes made from the wide-bandgap semiconductor SiC have exhibited an exceptional capability to switch and control large amounts of electrical power. It has emerged as an important semiconductor for power electronics in the high voltage regimes such as for automotive applications in the ~ 1000 V range. Gallium nitride

HEMTs offer exceptional high frequency switching performance in the ~ 100 V voltage window. These two semiconductors of bandgaps in excess of 3 eV are therefore slowly replacing the incumbent technologies in many high-voltage applications by significantly enhancing their energy efficiency and shrinking their size and weight for new generations of compact power sources and convertors.

Light-emitting diodes and lasers based on III-V semiconductors use nanometer-thick quantum wells in their active regions for efficient light generation (**Chapter 29**). GaAs and InP based lasers, modulators, optical amplifiers, and photodetectors have enabled optical communication systems with high bandwidths with the capacity to multiplex several high data-rate channels. VCSELs based on III-V semiconductors are embedded in mobile phones for face recognition. Quantum-dot and quantum-well based III-V lasers are also being integrated with silicon to form photonic integrated circuits. Intersubband and interband quantum-cascade lasers and infrared photodetectors are ushering in new regimes of long-wavelength imaging applications with applications in environmental monitoring by the detection of pollutants in the atmosphere.

As a quantum facelift to a 100-year old technology, GaN based LEDs are replacing incandescent light bulbs (Fig. 30.2). The energy saving by moving to semiconductor lighting is substantial. The energy efficiency and portability of LEDs is bringing lighting to places that lack established infrastructure. Lighting is simultaneously ushering in new forms factors and applications, such as backlighting of display screens and in automotive and industrial lighting applications. MicroLED technology is expected to continue this change well into the next decade.

Unlike incandescent bulbs, LEDs are designed at the quantum mechanical level to efficiently convert electrical energy to light. Therefore they can be modulated at high speeds. This is leading to new potential applications such as in data transmission using visible lighting, called visual light communication (VLC), or Li-Fi.

LEDs made from organic semiconductors (OLEDs) offer vibrant large area displays that are currently used in mobile phones and TV screens. Semiconductors such as indium gallium zinc oxide (IGZO) deposited over large areas offer carrier mobilities of ~ 10 cm^2/V.s and are used today in thin-film-transistor (TFT) screens for low power dissipation.

Solar cell efficiencies (**Chapter 28**) are climbing by refinement of established semiconductors such as amorphous and crystalline silicon, GaAs, and multijunction tandem cells, and the cost in \$/Watt is dropping continuously. Several new semiconductor materials ranging from organics to hybrid perovskites are being investigated for cost-effective large area solar cell panels in the future, but their efficiencies need to improve to be competitive.

Atomically thin semiconductors (**Chapter 11**) such as graphene and transition metal dichalcogenides are being explored for potential new applications stemming from their extreme thinness, and new physical

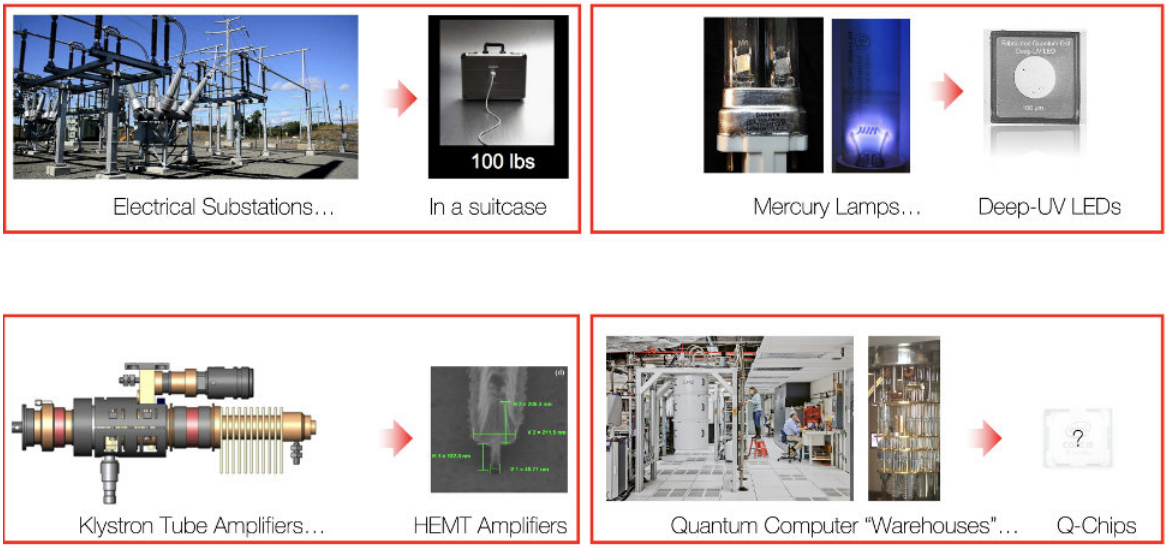


Fig. 30.3 New forms of semiconductors are expected to enable reductions in size in several fields, just as it did for computation and lighting in the past. These fields range from high-voltage and power electronics to ultra-high frequency RF electronics to deep-UV photonics, to quantum computation.

properties that they exhibit.

Very narrow bandgap semiconductors such as HgTe (which has a *negative* energy bandgap) and its alloys with CdTe have been used for very long-wavelength infrared photodetectors for night-vision and IR imaging. These mercury-cadmium-telluride (MCT) detectors enable imaging in the $10 - 20$ μm spectral window and are typically cooled to reduce the noise generated by thermal carriers (**Chapter** 28).

The electronics and photonics applications of semiconductors are being continually expanded by the control and understanding of wider energy bandgap semiconductors. For example, semiconductor devices of bandgap close to 6 eV are now becoming possible using aluminum nitride (AlN). This is enabling photonics in the deep-UV wavelength range using semiconductors, which had remained unattainable for decades. Deep-UV LEDs and lasers are expected to make a far deeper connection to biology than semiconductors have done in the past, because after nearly 80 years of semiconductor research and development, now it is finally possible to directly cause electronic transitions in biological molecules using semiconductor light sources. The realization of infrared semiconductor lasers revolutionized communication, and visible lasers and LEDs revolutionized solid-state lighting. For the very first time in history, semiconductor light sources are emitting deep-ultraviolet (DUV) photons of wavelengths 200–300 nm. The energies of these photons are large enough to directly interact with biomaterials such as cells, DNA/RNA, and proteins. The time is ripe for the convergence of semiconductor DUV photonics with biosystems.

Ultra-wide bandgap semiconductors such as AlN, BN, diamond and

Fig. 30.4 The "Traitorous Eight", who left Bill Shockley's semiconductor company and formed Fairchild Semiconductors to make silicon transistors in the San Francisco bay area, giving birth to the silicon valley. See if you can spot Robert Noyce, the inventor of the integrated circuit (IC), and Gordon Moore, known for Moore's law. A sub group later started Intel.

gallium oxide are also expected to bring about advances in electronics. These materials will enable power electronic switches that will significantly reduce electrical power conversion losses of today, operate at high temperatures, and their intrinsic radiation-resistance makes them suitable for future space and oil and gas exploration. Together with GaN and SiC, the ultra-wide bandgap semiconductors are foreseen to play a major role in the making of a more energy efficient, distributed, and resilient electrical power grid of the future, with the dexterity and efficiency which semiconductor devices are known to bring about when they replace bulkier components with mechanical moving parts.

30.2 Looking back and forward

One may say that the ubiquitous transistor is a **phase-transition** device because with the gate voltage, we move the Fermi level of the conducting channel from inside the bandgap (insulating phase) into the band (metallic phase). In every on/off switching event, the gate voltage causes a metal-insulator transition. This is correct, and indeed is the reason why the device has been the most transformative in the history of solid-state electronics. But in the future are there more phase transitions we could potentially exploit with semiconductors?

It is helpful to view the field of semiconductor materials and devices through the lens of **electronic phases** and **phase transitions** between them. Though this concept is prevalent (perhaps even the *defining* one) in fields such as materials science (e.g. crystal or structural phases) and condensed matter physics (e.g. ferromagnetism and superconductivity as various phases of the electron charge and spin), the field of semiconductors has not adopted or exploited this point of view to date, at least not to the level as others have.

There is a good reason for this state of affairs. The typical semiconductors covered in this book and mentioned in the last section have well-defined energy bandgaps (**Chapters** 8–13), and are doped to obtain n-type or p-type conductivity (**Chapters** 14, 15). Heterostructures and quantized geometries are used to control the potential and kinetic energies of electron and holes that flow in them. With this simple set of rules[1] and building blocks (**Chapter** 16), semiconductor scientists have made wonderful devices ranging from diodes to transistors, and from solar cells to LEDs and lasers.

But with progress in the materials science and condensed matter physics of semiconductors it has become clear that the Venn diagram of semiconductor electronic phases has rich intersections with other electronic phases. While standard semiconductor electronic and photonic devices are based on single-particle physics and phases, materials such as ferromagnetics, ferroelectrics, and superconductors exhibit **correlated electron phases** and **phase transitions**. This is rather exciting for the field, one that can yield rich rewards if pursued further. We can create semiconductors that are simultaneously ferroelectric, superconducting, ferromagnetic (or antiferromagnetic), and thereby sample a rich range of electronic phases inaccessible in traditional semiconductor materials. These are certainly not limited to electronic phases: hybrids of electrons and phonons in strongly coupled semiconductor acoustoelectric structures, or of electrons and photons in strongly coupled light-matter hybrids enrich the phases accessible in semiconductors. A range of new device applications can emerge from the new classes of hybrid semiconductors that exploit phase transitions.

As a few examples, the 2D semiconductor material indium selenide is also ferroelectric, as are wider bandgap semiconductors such as various oxides and AlScN. Several semiconductors such as diamond, MoS_2, and even graphene exhibit rich superconducting phases. Unlike in traditional materials, in semiconductors the ferroelectricity and superconductivity phases can be modulated with field effect, carrier injection, or with light! These correlated electronic phases of matter offer opportunities to remove the traditional ”walls” that exist today between memory, logic, and communication systems. With cloud computing and a data-center based computing future, semiconductor enthusiasts need not limit their dreams to devices that operate at room temperature: cryogenic operation (e.g. keeping the electronics in a data center immersed in liquid helium) is feasible if there is a justified boost in functionality and performance.

The next few sections discuss some of these effects. I must alert the reader that many of these areas are not as mature as what was discussed earlier in the book, and are subjects of ongoing research.

30.3 Ferroelectric semiconductors

Semiconductors of the hexagonal crystal structure such as GaN, InN, AlN and ZnO are strongly piezoelectric and also possess spontaneous polarization, but are ordinarily not ferroelectric. Ferroelectrics[2] are

[1] It is fair to say that the four properties of an electron: the charge, the mass, its momentum, and its spin – are all at the mercy of the designer of the semiconductor crystal or heterostructure.

[2] Ferroelectricity was discovered in 1920 by J. Valasek in Rochelle salt.

the electric-analogs of ferromagnets, with the electric dipole moment playing a role analogous to the spin polarization of magnets. There are several mechanisms of ferroelectricity – most of which originate in minute structural distortions of the location of atoms in the crystal that modify the net electric polarization. Switching of ferroelectrics occurs by the nucleation and expansion of domains, as explained in the Landau theory of phase-transitions.

The defining feature of ferroelectrics is that the polarization can be switched by an external electric field, just like an external magnetic field can switch the polarity of a ferromagnet. Thin films of ferroelectric materials such as lead zirconium titanate (PZT) have been used for ferroelectric random access memories (FeRAMs). There is significant recent interest to integrate ferroelectric materials into the gate dielectric of transistors directly to *merge* memory elements with logic, for non-volatile logic[3].

Some semiconductors are being discovered to be ferroelectrics too. Examples include the layered semiconductor In_2Se_3 and the wide-bandgap material AlScN. In fact several well-known ferroelectrics such as perovskite oxides (e.g. $BaTiO_3$) are wide bandgap semiconductors that could potentially also be doped and imparted semiconducting properties. Ferroelectricity is also found to emerge in unconventional structures such as multilayers of 2D crystals of graphene and hexagonal boron nitride (hBN) dependent on the stacking order and layer number. When ferroelectric semiconductors are used as the active channels of semiconductor electronic and photonic devices with doping and heterostructure control, several new functionalities are expected to emerge. The wide tunability of the optical properties and the capacity to pole ferroelectrics can enable several new methods to mold the flow of light or microwaves on integrated circuits, with applications in RF/mm-wave electronic, photonics, and quantum computation and communications.

30.4 Ferromagnetic semiconductors

Magnetic metals have been integrated with semiconductors, and spin-polarized carriers have been injected into semiconductors from magnetic contacts. Typical semiconductors are not ferromagnetic: therefore they allow for spin-polarized carriers to exist over length scales limited by spin-dephasing scattering events.

Replacing the metal cations of II-VI and III-V semiconductors with transition metal atoms such as Fe or Mn introduces spin-polarized electrons near the band edge. If the spin polarized carriers are mobile, they can magnetize the entire semiconductor crystal. Such **dilute magnetic semiconductors** exhibit magnetic order which can be turned on or off by modulating the carrier density with gate voltages or carrier injection. Thus they exhibit tunable magnetism.

GaMnAs is a dilute magnetic semiconductor where the Mn atomic

[3] Closely linked to ferroelectric order is antiferroelectric ordering, with potential usage in non-volatile logic.

concentration is typically 5–10 %. The Curie temperature of GaMnAs and its alloys or quantum well structures with In can today exceed 100 K, but still falls short of 300 K. It is foreseeable that in the future semiconductors in which magnetism persists above 300 K will be found. Like the ferroelectric order, such magnetic semiconductors can also offer pathways to merge logic and memory functionalities, and new applications can be expected to emerge from them. Recently magnetic ordering has been observed to emerge in stacked layers of 2D semiconductors, which can also bring ferromagnetic ordering into semiconductor families.

30.5 Multiferroic semiconductors

Insulating perovskite crystals such as $BiFeO_3$ and $LuFeO_3$ exhibit electronic and magnetic polarization simultaneously, i.e., they are ferroelectric *and* ferromagnetic (or ferrimagnetic). What makes such materials especially interesting is that the two phases: electronic polarization and magnetic polarization, are *coupled*. This coupling means that, for example, by switching the ferroelectric polarity by field-effect, one can also switch the magnetic polarity. This mechanism is rather remarkable because it is a direct field-effect control of magnetism, which can allow for new forms of memory and logic devices in the future. There are very few known multiferroic materials, and their semiconducting properties are poorly understood today.

30.6 Superconducting semiconductors

Several semiconductors (e.g. Si, Ge, diamond and others), when doped heavily beyond the Mott criteria into the metallic regime, exhibit superconductivity at low temperatures. The typical critical temperatures fall below $\sim$ 10 K. Superconductivity is also observed in transition metal dichalcogenide semiconductors such as MoS_2 where it is created by field effect by populating the band with a large density of free carriers. A new discovery is that of superconductivity in twisted bilayer graphene, in which at certain "magic-angles" of the relative twist angle between the layers the density of states near the Dirac point diverges due to formation of Moiré superlattices in the bands.

Heterostructures of semiconductors and superconductors have successfully been used to inject superconducting Cooper pairs from the superconductor into semiconductor channels. Josephson junctions and simple circuits of this kind have the potential to exploit the macroscopic quantum coherence of superconductors and the modulation capabilities of semiconductors. The large mismatch in the energy bandgaps of semiconductors (typically $\sim$ 1 eV) and the gaps of superconductors (typically $\sim$ 1 meV) typically poses an impediment on a direct integration of the functionalities of the two electronic phases.

Heterostructure of superconductors with semiconductors are of significant future interest for the areas of energy efficient computing and of quantum computation and quantum communications. Furthermore, since a significant amount of computation in the future will be performed in data centers, the cooling restriction may not be as prohibitive as it is todat for personal computers and hand held devices.

30.7 Semiconductors for quantum communications

Communication systems in the past used to be powered by moving information encoded in electrons in wires. Today they are increasingly replaced by photons of optical wavelengths moving either in optical fibers (e.g. light of $\lambda = 1.55$ μm), or by micro (and mm-)wave photons (waves of $f = 1 - 100$ GHz) moving in free space. These waves are generated on the transmit side by semiconductor lasers for optical wavelengths, and by transistor amplifiers + antennas for microwave frequencies, and are detected by semiconductor photodetectors or low-noise amplifers on the receive side of communication links. Optical or microwave amplifiers that provide gain are inserted in the communication channel too to boost the strength of the signals as they propagate and encounter loss.

The classical communication systems of today therefore are actually based on semiconductor devices that depend on quantum mechanics for their operation. The number of photons that carry a zero or one in digital communications is very large, several orders larger than unity. An attribute of classical communication systems is that they can be *eavesdropped*: the information flowing through the channel can be read off without the sender or the receiver knowing this has happened. This is tied ultimately to the large number of photons carrying the information.

It is now realized that if the quantum properties of either single or very few photons carrying the information is exploited, then **quantum communication systems** can be made more **secure**. Cryptography, the game of codemaking and codebreaking, appeals uniquely to the competitive human spirit. It is a measure of our intellect and ingenuity: what one person can cleverly hide, the other can surely seek. In the past the use of the art, science, and engineering of cryptography was limited to the military, and to national and international diplomacy. It did not directly impact a large part of the population. But today, the everyday personal and financial transactions of a vast and growing population on the earth is no more performed by in-person meetings, exchange of secret letters, or by the exchange of paper money. The personal and financial information is converted directly from the brain of a person to electronic 0 and 1 bits when he or she types, or speaks into a computer. The digital 0 or 1 bits are stored, and shuttled around by electrons on integrated circuit chips, or by optical pulses, between chips across computers, jumping between cities, states, and countries at the speed of light. Most modern online electronic financial transactions are protected by codes such as the RSA, which can, in principle, be broken by either a supremely faster classical computer today, or by the discovery of a new mathematical theorem tomorrow.

The only unbreakable code, called the *one-time pad* cipher (and its variants) is a scheme of secure communication in which the information to be exchanged is first encoded using a key which is a string of random numbers. The encrypted message is transmitted over a public channel from Alice to Bob. An eavesdropper Eve can intercept it, but it is impossible to decode it without the key. This one-time pad cipher was mathematically proven to be unbreakable by Claude Shannon, as long as the coded message is shorter than the key used to code it. But the catch is that the key must be stored, and transmitted in complete secrecy, and known only to Alice and Bob. The implementation of this scheme thus requires the secure physical storage and exchange of a key – which of course can be (and has been, several times in history) intercepted or cloned by Eve without the knowledge of Alice and Bob. This is because any system that follows the laws of classical mechanics allows for unfettered observation without affecting the state of the system.

In the face of this uncertainty, can one *guarantee* the security of transmitted data against technological, or mathematical superiority of an eavesdropper – both now, and forever in the future? In the 1990s, Bennett, Brassard, and Ekert, the pioneers of quantum cryptography showed that the answer is yes. It is possible as long as three pillars of the deepest physical laws of quantum mechanics stand firm: 1) the quantum no-cloning theorem, 2) the impossibility of the measurement of a *single* quantum state without collapsing the state (e.g. impossibility of the measurement of the polarization of a single photon), and 3) quantum entanglement. These physical principles have withstood theoretical and experimental challenges for nearly a century, since the conception of quantum mechanics itself. As long as these laws hold, a scheme that guarantees security of communication was developed by Bennett and Brassard in 1984, and named the **BB84** scheme. Ekert further exploited quantum entanglement to devise a scheme that guarantees not just completely secure transmission, but also completely secure quantum storage.

In the 1980s, the first atomic sources of single photons became experimentally accessible, and single photon antibunching was experimentally observed. If a single photon wavepacket is split by a beam-splitter, the single photon detectors in the two channels will record a co-incidence, because if the single photon is detected in one channel, it *cannot* be detected in the other – which is the fundamental difference between quantum and classical description of light. On the other hand, classically-produced photons occur in multiples – multiple photons can enter the beam-splitter at the same time. If so, it is possible to measure a photon on both detectors. Therefore, a measurement of the

2-photon correlation function of the two detectors, $g^{(2)}(\tau)$, provides an experimental indication of the classical or quantum nature of the photon source. A characteristic decrease in $g^{(2)}(\tau = 0)$, known as antibunching, tells us that we have a quantum source, whereas if no such dip is present, then the source is classical.

The quantum-secure BB84 communication scheme was experimentally implemented using classical weak light pulses from green LEDs at IBM, by exploiting the polarization of single photons. The BB84 scheme encodes each bit in a single polarized photon. This scheme has come a long way since the initial demonstrations: the secure communication of the Swiss election data in 2007 was one of the first mass demonstrations of quantum communications and captured the interest of the population. Single photon sources may be purchased today from startup companies (such as IDQ, https://www.idquantique.com). However, most of these single photon sources are based on generating extremely weak light pulses by attenuating the classical light emitted from a semiconductor LED or a laser, like the original IBM demonstration. Though classical light sources can be used to produce on-demand single photons, they suffer from several drawbacks that limit their widespread use.

Ideally, single photon emitters for secure quantum communication should have three key properties. First, they should be bright, ideally with emission *into the detection system* of 100s of megahertz to gigahertz emission rates. Second, they should be *single* - such that one and only one photon be emitted into a single time-bin. Third, they should be capable of encoding a quantum state, which can be done using either polarization, frequency, or time-bin, for example.

A key difference between using isolated quantum states as compared to "decoy states" (e.g. attenuated lasers), is that an isolated quantum emitter can only emit photons one at a time – they are truly single. Indeed, if multiple photons appear in a time bin, then quantum communication is not secure to eavesdropping. This, therefore, limits the *rate* of current commercial quantum key distribution systems that are based on decoy states. A weak classical light pulse $|\alpha\rangle$ can be considered an eigenstate of the photon annihilation operator $\hat{a}|\alpha\rangle = \alpha|\alpha\rangle$. In particular, the photon number state of a weak classical light source from a laser or a LED follows the Poisson distribution $|\alpha\rangle = e^{-\frac{\mu}{2}} \sum_{n=0}^{\infty} \alpha^n / \sqrt{n!} |n\rangle$. Here $\mu = |\alpha|^2$ is the average photon number. In order to make it sufficiently improbable to detect two or more photons at once, it is necessary to use $\mu \sim 0.2$, which means that most of the time, there is no photon. One must not forget that nonclassical light emission can be obtained by careful quantum engineering of the semiconductor heterostructures and transport of injected electrons and holes. Typically such light sources fall somewhere between the classical and true single photon sources. In contrast, a single quantum system can physically only emit only one photon at a time (e.g. without spurious background photons, $|\alpha\rangle = |1\rangle$). If the single photon emitter is operated near saturation with high-efficiency optical collection, it is possible to get a single photon at every time step.

30.8 Semiconductors for quantum computation

Digital computation today is achieved by performing logic operations following the von Neumann architecture. The energy of a collection of electrons is required to be in one of two distinguishable states: a logic high (= 1) or a logic low (= 0). Logic operations with 0's and 1's, or Boolean logic is performed to create information when gates in the form of transistors and their combinations switch the states between 1 and 0. The results are stored as strings of 1's and 0's. In classical or digital computation, each collection of electrons sample two points in a Bloch sphere (**Chapter 20**), though the underlying physics of actual manipulation and storage of data today is inherently quantum mechanical. In other words, classical computation can *in principle* be performed with larger mechanical objects that follow the rules of classical mechanics. In that case it sure would be far less energy efficient, bulkier, and slower than if the computation is done by electrons following the rules of quantum mechanics. But there is no fundamental hurdle in actually performing the computation. The analogy is that a simple calculation such as the multiplication of two large numbers can be done on either a semiconductor microprocessor, or with an abacus.

A two-level quantum system can sample the entire surface of the Bloch sphere instead of just two points (**Chapter 20**) by exploiting the superposition principle, which is embedded in the *phase* of the wavefunction. This leads to a vast increase in the number of states that a small number of quantum particles can access and exploit for computation. This form of computation exploits the quantum phase (and not just the amplitude of the wavefunction), and the principles of superposition and entanglement. Since this is not possible in classical computation, it is called **quantum computation**. The two-level quantum system is called a quantum-bit or **qubit** as an explicit acknowledgement of its classical pedigree. This vastly enhanced space to perform computation has led to the proposal of innovative algorithms that promise to solve certain classes of problems that are very difficult for classical computers. This class of ″NP-hard″ problems require exponential time scales with classical computers when the problem complexity scales, but by the use of entanglement of quantum states, can potentially be solved in polynomial time with quantum computers.

Multiple experimental methods are being explored today to realize qubits. The basic ingredients (**Chapter 20**) necessary are two-level systems with an external means to cause transitions between them, and a high degree of coherence such that once left alone, the quantum state can evolve according to the time-dependent Schrödinger equation and sample the large number of states. These seemingly simple requirements pose a difficult technological challenge, one that is being addressed by exploring different experimental embodiments of the qubit.

A modern example is the superconducting microwave qubit. For such qubits collective superconducting states of a Josephson junction are used as the two-level system, and a microwave photon causes transitions between them. Superconducting transmission lines form a high-Q cavity so that the coherence is maintained over long periods of time. The number of such qubits are currently up to ~ 100, far smaller than what is needed to perform meaningful quantum computation. The current challenge is one of maintaining coherence over long time scales so that more qubits are entangled to reap the benefit of quantum computation. The source and detection of the microwave photons in this qubit scheme are semiconductor transistor amplifiers based on III-V semiconductors. The generation of microwave pulses needed for quantum computation also require precise microwave circuits based on semiconductors.

Though semiconductors play a role in superconducting qubits, another qubit approach that promises scaling of qubits to far smaller length scales puts semiconductors at the heart of the action. These use lithographically defined semiconductor quantum dots. The spin-state of electrons in these quantum dots that are formed in the bands of semiconductors (such as silicon or SiGe alloys) are the two-level system. Such **semiconductor spin-qubits** place stringent requirements on the precision of lithography and fabrication, and of interfacing the qubit states with photons of appropriate energy.

Other qubit approaches use the two energy levels of defect states (e.g. vacancies) in semiconductors, or of specially selected atoms or ions that are embedded in the semiconductor matrix. Special care is taken to ensure that the two energy levels do not fluctuate in time due to lattice vibrations or other means of coupling to the crystal lattice. Examples of this kind are nitrogen vacancy centers (**NV-centers in diamond**), or Rb ions implanted into silicon or SiO_2. A major challenge today is the deterministic placement of these defects in the crystal. Once placed, various waveguiding and photonic integration schemes are used to bring photons into strong interaction with the defects, and to entangle multiple qubits using light.

Though photons of microwave frequency in the ~ 10 GHz range have the advantage of a high level of circuit integration for both superconducting qubits and semiconductor spin qubits, they are easily destroyed by room temperature thermal energy. This is because for them $\hbar\omega \leq k_b T$ for $T = 300$ K, and thermal vibrations of atoms of molecules are enough to absorb or emit microwave photons. On the other hand visible photons have $\hbar\omega >> k_b T$ and are therefore far more robust to room temperature thermal fluctuations. Schemes are currently being explored to exploit the advantages of both frequency regimes in qubit architectures that use both microwave and optical photons with conversions between them.

Quantum computation is in a nascent stage. It will be exciting to see which path it takes, and what role semiconductors will play in it in the near future. The level of precision necessary to make qubits and to achieve quantum communications is already yielding ancillary dividends. As an example, the exquisite sensitivity of NV centers in diamond to magnetic fields is now used to fabricate new generations of precision magnetometers.

30.9 Semiconductors for energy

Semiconductors remain at the forefront of the clean energy generation through their use in solar cells (**Chapter 28**). Innovations in semiconductor materials and solar cell device design, and in the fabrication, assembly, and installation of solar cells continue to push the adoption of solar energy as a genuine investment for the future of the planet.

Semiconductor thermoelectrics (**Chapter 21**) perform appreciably, but their limited ZT values lead to lower efficiencies of energy conversion than solar cells. So their primary applications are not for generation of energy, but for fast electronic control of temperature. But in applications where light energy is not available, especially in space, thermoelectric energy generation is very useful as discussed in a later section of this chapter.

The replacement of incandescent light bulbs by white LEDs (**Chapter 29**) has led to significant energy savings owing to their high efficiency. The impact has been substantial in indoor residential lighting as well as in industrial, business, and automotive lighting applications. The ability to tailor the emission spectrum via the design of semiconductor quantum wells and by the choice of phosphors for photon energy downconversion is leading to new and innovative applications of LEDs. These range from lighting that adapts to the environment, to efficient indoor and urban farming and agriculture as described in the next section.

The rapid automation by ”electrification” or digitization of data management, banking, government functions, and commerce means that vast amounts of energy now flows into and out of transistors and memories. Therefore, the energy efficiency of each transistor and memory element and the communication links that move the data from where it is stored to where it is used is expected to play an increasingly important role in the future, at par with the need to increase the efficiency of solar cells.

One of the largest areas of energy efficiency to which semiconductors are bringing their characteristic miniaturization, speed, efficiency, and reliability today is in power electronics. High-voltage electronic switches and diodes manipulate and condition electric energy in a wide variety of electrical appliances and equipment used in residential and business establishments. Electric power delivery and management in electric vehicles and transportation is highly transformative and is a large area of application of wide-bandgap semiconductors for efficient power switching. Significant energy savings are already made possible by the use of energy-efficient high voltage semiconductor switches in passenger trains.

Ultra-wide bandgap semiconductors such as AlN and diamond are expected to enable energy efficient power switching and energy conversion at even higher voltages than what SiC and GaN are capable of today. They will therefore enable increased efficiency, resilience, and agility of future electrical grid infrastructure and make them robust to outages due to weather and accidents. Their light weight and compact nature can potentially enable innovations in all-electric flight, which remains prohibitive today. Of course this also requires significant innovations in lightweight and high energy-density electric energy storage technologies. These ultra-wide bandgap semiconductors are currently at the level of maturity that SiC was in the 1990s or GaN in the early 2000s, and several aspects of doping, defects, interfaces, and heterostructure formation must be understood for such applications to turn into reality.

Ultimately the highest voltages semiconductors can sustain while allowing the functions of switching and gain are related to the high-energy phenomena of avalanche breakdown and tunneling (**Chapter 24**). Several of these phenomena have counterparts not in condensed matter but in high-energy physics (e.g. impact ionization in semiconductors is analogous to pair production in high-energy physics). As ultra-wide bandgap semiconductors are taken to energy densities that traditional semiconductors have not been subjected to in the past, it is possible that currently unanticipated phenomena are discovered, and the field of semiconductor physics itself is enriched!

30.10 Semiconductors for healthcare and agriculture

The use of semiconductor LEDs with tailored optical output spectrum enables photosynthesis in plants. In turn this is driving an increasing interest in indoor and urban horticulture and farming of vegetables and food items. Indoor farming can be year-round, and can be immune to inclement weather and climate change. The ability to tailor the spectra also provides the ability to control photosynthesis (blue and red wavelengths), longer wavelength red for growth and flowering, and even UV to prevent infection. This area is in its infancy today: semiconductor light sources will find increasing use in it in the future as the need to feed and sustain the growing world population in a sustainable and energy-efficient manner increases.

Semiconductors already play a large role in healthcare by being the electronic brain and the device components for imaging in almost all diagnostic and therapeutic equipment ranging from MRIs to those use in surgery. Digitization of medical data and availability of the genome is allowing the introduction of computational and algorithmic approaches to healthcare and preventive treatment. Several implants such as pacemakers and those for visually impaired use lightweight, rugged and compact electronic components only feasible with semiconductors. UV LEDs are finding increasing use in curing and for disinfection of hospital equipment. Semiconductor based micro-electromechanical equipment are enabling new forms of precision mechanical actuation and sensing in confined spaces to enable, for example, measurement of blood pressure, rapid analysis of samples via lab-on-a-chip methods, and in precision drug delivery.

30.11 Semiconductors for space exploration

Semiconductors have been instrumental in space exploration missions by providing lightweight, energy-efficient, reliable, and radiation resistant components for energy generation (solar cells), electronic control (transistor circuits and microprocessors), imaging (semiconductor photodetectors), Hall sensors, and electronic components for high-voltage relays and power convertors. Increasing use is found for semiconductor lasers for positioning, spectroscopy, and communication links and data transfer, and also for detection of atmospheric gases and elements. The semiconductor devices and components used in space applications have to be resistant to high-energy radiation.

For space missions to extreme environments such as near the Sun or high temperature planets, wide-bandgap semiconductors such as SiC, GaN, AlN, and diamond are being actively investigated and deployed because low interband thermal carrier generation preserves their semiconducting nature. In fact some of these devices paradoxically perform *better* at high temperatures when deep acceptor or donors are thermally ionized making them more conductive when hotter! Semiconductor technology is expected to play an increasingly important role in the use of satellites and in future manned missions to other planets.

Semiconductors have made it possible to reach the farthest distances from the earth: the Voyager is powered by silicon-germanium (SiGe) semiconductor thermoelectric generators which convert the radiation from nuclear reactions in radioisotopes into electrical energy. In a way, they are carrying their own "Sun". These human-made semiconductor devices that have gone into interstellar space! They carried 64 kb of memory. The two Voyager probes were launched in 1977, and are still transmitting data to the earth today (in 2022).

30.12 Social impact of semiconductors

We end this chapter and the book with a brief discussion of the social impact of semiconductors. The first direct impact is through the economics the semiconductor industry has created by the technologies discussed in this chapter and earlier in the book. But the impact of semiconductors on human psyche, relationships, and social interactions and networks is equally transformative.

Semiconductor technologies are causing a major upheaval via automation, by making a large number of tasks cheaper and efficient. This indirectly implies a transformation of several industries and jobs that used to employ large fractions of the population. Nations and societies that accept that this change is inevitable, and put policies in place that account for this change stand to benefit from it.

This implies that education and training should place increasing emphasis on creativity and the ability to cleverly use the rather powerful electronic infrastructure that is being created.

It is misleading to think that human beings control computers, information, and the World Wide Web. It is not a one-way traffic. The electronic network also today controls several of our actions and even our beliefs. It gives us unfettered access to our loved ones by giving instant connectivity to family and friends across the globe, and also helps find and reconnect with long-lost friends and acquaintances. The same unfettered access also makes the electronic network a powerful tool to influence and ”control” our opinions and beliefs, and ultimately our actions. It is anticipated that the current era of hyper-connectivity will go through several growing pains as we collectively learn how to use it in a balanced manner.

But the future of semiconductors and what they can do and will do is bright, and far more exciting that what we have seen till today (see the quote from Marie Curie in the Preface of this book). There will be a role for semiconductors in every aspect of human endeavor, some of which were discussed here. To share the future excitement, we end this book in the same spirit as we began – by saying: **let the next semiconductor adventure begin!**

30.13 Chapter summary section

In this chapter, we discussed:

- The latest technological applications of semiconductor materials and devices,
- New areas that the field of semiconductors are moving into, and are expected to impact in the near future, and
- How semiconductors today have enormous and increasing impact on energy, healthcare, agriculture, transportation, and on the society.

Further reading

The Story of Semiconductors and **Semiconductors and the Information Revolution** by J. Orton provides a glimpse into some aspects of the topics discussed in this chapter. **Ferroelectric Semiconductors** by V. Fridkin has a treatment of the new device physics to be expected in semiconductors that simultaneously exhibit ferroelectricity. **The Microchip Revolution** by L. Oliver and E. Wilder takes the reader through the history of how the microchip was possible. Surprisingly, there are no books written yet on the social impact of semiconductors – perhaps because we are living in it right now!

Exercises

(30.1) **A short parting essay**
The materials covered in this chapter will evolve rapidly in time. How innovative the readers of this book are will be reflected in how soon several sections of this chapter will be outdated! Therefore the only exercise here is to re-assess each section of this chapter and to write a short update them to the most current developments.

Appendix

A

A.1 What is in the appendix? 847
A.2 Semiconductor Formulae 847
A.3 Physical properties of 3D and 2D semiconductors 847
A.4 References for the appendix 852
A.5 Physical constants 853

A.1 What is in the appendix?

This appendix contains the following:

- Fig. A.1 provides a summary of the important formulae that govern the quantum physics of semiconductor materials and devices. The reader can use this as a look up table for expressions for density of states, carrier densities, ballistic transistors, tunneling, or photonic processes such as stimulated emission spectrum and optical gain spectrum. The expressions are tabulated for different dimensionality of quantum confinement, and generalized expressions for d dimensions highlights the universality of the formulae.
- Table A.1 summarizes the properties of selected semiconductors that are used in several technological applications today. Table A.2 presents the properties of a more exhaustive list of 3D semiconductors. Table A.3 lists properties of 2D semiconductors. Since 2D semiconductors are less exhaustively studied, several vacant table entries are expected to be filled in the future.
- The data in Tables A.1, A.2, and A.3 were collected from a list of texts, websites, and journal papers[1]. These sources are listed in the references section of this appendix. The references are provided in the tables with the corresponding semiconductors. A large chunk of the references are for semiconductors such as monolayer 2D materials, and β Ga_2O_3 and related ultra wide bandgap semiconductors, whose properties are currently under active exploration. These values should be used with care, and updated with newer experimental reports wherever appropriate.
- Table A.4 lists the numerical values of the fundamental constants that appear in this book.

A.2 Semiconductor Formulae

A.3 Physical properties of 3D and 2D semiconductors

---

[1] These three tables were assembled with generous help from Dr. Reet Chaudhuri, and Yifan (Frank) Zhang, both from Cornell University.

	0 Dimension	1 Dimension	2 Dimensions	3 Dimensions	d Dimensions
Conduction Bandstructure	E_c	$E_c + \frac{\hbar^2 k_x^2}{2m_c^\star}$	$E_c + \frac{\hbar^2}{2m_c^\star}(k_x^2 + k_y^2)$	$E_c + \frac{\hbar^2}{2m_c^\star}(k_x^2 + k_y^2 + k_z^2)$	$E_c + \frac{\hbar^2}{2m_c^\star}(\sum_{i=1}^{d} k_i^2)$
Conduction Band DOS $g_c^d(E)$	$g_s g_v \delta(E - E_{cn})$	$\frac{g_s g_v}{2\pi}(\frac{2m_c^\star}{\hbar^2})^{\frac{1}{2}} \frac{1}{\sqrt{E-E_c}}$	$\frac{g_s g_v}{2\pi}(\frac{2m_c^\star}{\hbar^2})\theta(E - E_c)$	$\frac{g_s g_v}{(2\pi)^2}(\frac{2m_c^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{E - E_c}$	$\frac{g_s g_v}{(4\pi)^{\frac{d}{2}}\Gamma(\frac{d}{2})}(\frac{2m_c^\star}{\hbar^2})^{\frac{d}{2}}(E - E_c)^{\frac{d-2}{2}}$
Conduction Band-Edge Effective DOS N_c^d	–	$g_s g_v(\frac{2\pi m_c^\star k_b T}{h^2})^{\frac{1}{2}}$	$g_s g_v(\frac{2\pi m_c^\star k_b T}{h^2})$	$g_s g_v(\frac{2\pi m_c^\star k_b T}{h^2})^{\frac{3}{2}}$	$g_s g_v(\frac{2\pi m_c^\star k_b T}{h^2})^{\frac{d}{2}}$
Electron Density n_d	–	$N_c^{1d} F_{-\frac{1}{2}}(\frac{E_F - E_c}{k_b T})$	$N_c^{2d} F_0(\frac{E_F - E_c}{k_b T})$	$N_c^{3d} F_{+\frac{1}{2}}(\frac{E_F - E_c}{k_b T})$	$N_c^{d} F_{\frac{d-2}{2}}(\frac{E_F - E_c}{k_b T})$
Source Fermi Level E_{Fs} under bias V	–	$n_{1d} = \frac{1}{2} N_c^{1d}[F_{-\frac{1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) + F_{-\frac{1}{2}}(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$	$n_{2d} = \frac{1}{2} N_c^{2d}[F_0(\frac{E_{Fs} - E_c}{k_b T}) + F_0(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$	$n_{3d} = \frac{1}{2} N_c^{3d}[F_{+\frac{1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) + F_{+\frac{1}{2}}(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$	$n_d = \frac{1}{2} N_c^{d}[F_{\frac{d-2}{2}}(\frac{E_{Fs} - E_c}{k_b T}) + F_{\frac{d-2}{2}}(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$
Ballistic Electron Current Density J_d at voltage V	–	$\frac{q^2}{h} \cdot N_c^{0d} \cdot (\frac{k_b T}{q}) \cdot [F_0(\frac{E_{Fs} - E_c}{k_b T}) - F_0(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$	$\frac{q^2}{h} \cdot \frac{1}{2} \cdot N_c^{1d} \cdot (\frac{k_b T}{q}) \cdot [F_{\frac{1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) - F_{\frac{1}{2}}(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$	$\frac{q^2}{h} \cdot (2\pi)^3 \cdot N_c^{2d} \cdot (\frac{k_b T}{q}) \cdot [F_1(\frac{E_{Fs} - E_c}{k_b T}) - F_1(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$	$\frac{q^2}{h} \cdot \eta \cdot N_c^{d-1} \cdot (\frac{k_b T}{q}) \cdot [F_{\frac{d-1}{2}}(\frac{E_{Fs} - E_c}{k_b T}) - F_{\frac{d-1}{2}}(\frac{(E_{Fs} - qV) - E_c}{k_b T})]$
Valence Bandstructure	E_v	$E_v - \frac{\hbar^2 k_x^2}{2m_v^\star}$	$E_v - \frac{\hbar^2}{2m_v^\star}(k_x^2 + k_y^2)$	$E_v - \frac{\hbar^2}{2m_v^\star}(k_x^2 + k_y^2 + k_z^2)$	$E_v - \frac{\hbar^2}{2m_v^\star}(\sum_{i=1}^{d} k_i^2)$
Valence Band DOS $g_v^d(E)$	$g_s g_v \delta(E - E_{vn})$	$\frac{g_s g_v}{2\pi}(\frac{2m_v^\star}{\hbar^2})^{\frac{1}{2}} \frac{1}{\sqrt{E_v-E}}$	$\frac{g_s g_v}{2\pi}(\frac{2m_v^\star}{\hbar^2})\theta(E_v - E)$	$\frac{g_s g_v}{(2\pi)^2}(\frac{2m_v^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{E_v - E}$	$\frac{g_s g_v}{(4\pi)^{\frac{d}{2}}\Gamma(\frac{d}{2})}(\frac{2m_v^\star}{\hbar^2})^{\frac{d}{2}}(E_v - E)^{\frac{d-2}{2}}$
Valence Band-Edge Effective DOS N_v^d	–	$g_s g_v(\frac{2\pi m_v^\star k_b T}{h^2})^{\frac{1}{2}}$	$g_s g_v(\frac{2\pi m_v^\star k_b T}{h^2})$	$g_s g_v(\frac{2\pi m_v^\star k_b T}{h^2})^{\frac{3}{2}}$	$g_s g_v(\frac{2\pi m_v^\star k_b T}{h^2})^{\frac{d}{2}}$
Hole Density p_d	–	$N_v^{1d} F_{-\frac{1}{2}}(\frac{E_v - E_F}{k_b T})$	$N_v^{2d} F_0(\frac{E_v - E_F}{k_b T})$	$N_v^{3d} F_{+\frac{1}{2}}(\frac{E_v - E_F}{k_b T})$	$N_v^{d} F_{\frac{d-2}{2}}(\frac{E_v - E_F}{k_b T})$
Source Fermi Level E_{Fs} under bias V	–	$p_{1d} = \frac{1}{2} N_v^{1d}[F_{-\frac{1}{2}}(\frac{E_v - E_{Fs}}{k_b T}) + F_{-\frac{1}{2}}(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$	$p_{2d} = \frac{1}{2} N_v^{2d}[F_0(\frac{E_v - E_{Fs}}{k_b T}) + F_0(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$	$p_{3d} = \frac{1}{2} N_v^{3d}[F_{\frac{1}{2}}(\frac{E_v - E_{Fs}}{k_b T}) + F_{\frac{1}{2}}(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$	$p_d = \frac{1}{2} N_v^{d}[F_{\frac{d-2}{2}}(\frac{E_v - E_{Fs}}{k_b T}) + F_{\frac{d-2}{2}}(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$
Ballistic Hole Current Density J_d	–	$\frac{q^2}{h} \cdot N_v^{0d} \cdot (\frac{k_b T}{q}) \cdot [F_0(\frac{E_v - E_{Fs}}{k_b T}) - F_0(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$	$\frac{q^2}{h} \cdot \frac{1}{2} \cdot N_v^{1d} \cdot (\frac{k_b T}{q}) \cdot [F_{\frac{1}{2}}(\frac{E_v - E_{Fs}}{k_b T}) - F_{\frac{1}{2}}(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$	$\frac{q^2}{h} \cdot (2\pi)^3 \cdot N_v^{2d} \cdot (\frac{k_b T}{q}) \cdot [F_1(\frac{E_v - E_{Fs}}{k_b T}) - F_1(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$	$\frac{q^2}{h} \cdot \eta \cdot N_v^{d-1} \cdot (\frac{k_b T}{q}) \cdot [F_{\frac{d-1}{2}}(\frac{E_v - E_{Fs}}{k_b T}) - F_{\frac{d-1}{2}}(\frac{E_v - (E_{Fs} - qV)}{k_b T})]$
Interband Tunneling Current Density J_d^T	–	$\frac{q^2}{h} \cdot g_s g_v \cdot e^{-\frac{F_0}{F}} \cdot \frac{k_b T}{q} \cdot \ln[\cosh^2[\frac{qV}{2k_b T}]]$	$\frac{q^2}{h} \cdot g_s g_v \cdot (\frac{m_v^\star E_g}{3\pi\hbar^2} \cdot \frac{F}{F_0})^{\frac{1}{2}} \cdot e^{-\frac{F_0}{F}} \cdot V \cdot [(1 - \frac{V_0}{2\sqrt{\pi}V}) \cdot$ $\mathrm{Erf}(\sqrt{\frac{V}{V_0}}) + \sqrt{\frac{V}{\pi V_0}} e^{-\frac{V}{V_0}}]$	$\frac{q^2}{h} \cdot g_s g_v \cdot (\frac{m_v^\star E_g}{3\pi\hbar^2} \cdot \frac{F}{F_0}) \cdot e^{-\frac{F_0}{F}} \cdot V \cdot [1 - \frac{V_0}{V}(1 - e^{-\frac{V}{V_0}})]$	$\approx \frac{q^2}{h} \cdot g_s g_v \cdot (\frac{m_v^\star E_g}{3\pi\hbar^2} \times \frac{F}{F_0})^{\frac{d-1}{2}} \cdot e^{-\frac{F_0}{F}} \cdot V$
Electron-Hole-Photon Relation	$E_{cn} - E_{vn} = h\nu$	$E_g + \frac{\hbar^2 k_x^2}{2m_r^\star} = h\nu, (E_g = E_c - E_v, \frac{1}{m_r^\star} = \frac{1}{m_c^\star} + \frac{1}{m_v^\star})$	$E_g + \frac{\hbar^2}{2m_r^\star}(k_x^2 + k_y^2) = h\nu$	$E_g + \frac{\hbar^2}{2m_r^\star}(k_x^2 + k_y^2 + k_z^2) = h\nu$	$E_g + \frac{\hbar^2}{2m_r^\star}(\sum_{i=1}^{d} k_i^2) = h\nu$
Optical Joint DOS $\rho_{jnt}(h\nu)$ in $\mathrm{eV} \cdot \mathrm{cm}^{-3}$	$\frac{g_s g_v}{L_x L_y L_z}\delta[h\nu - (E_{cn} - E_{vn})]$	$\frac{g_s g_v}{2\pi L_y L_z}(\frac{2m_r^\star}{\hbar^2})^{\frac{1}{2}} \frac{1}{\sqrt{h\nu-E_g}}$	$\frac{g_s g_v}{2\pi L_z}(\frac{2m_r^\star}{\hbar^2})\theta(h\nu - E_g)$	$\frac{g_s g_v}{(2\pi)^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{h\nu - E_g}$	$\frac{g_s g_v}{(4\pi)^{\frac{d}{2}} L^{3-d}\Gamma(\frac{d}{2})}(\frac{2m_r^\star}{\hbar^2})^{\frac{d}{2}}(h\nu - E_g)^{\frac{d-2}{2}}$
Electron Density in the Optically Active Region	–	$N_c^{1d} F_{-\frac{1}{2}}(\frac{F_n - E_c}{k_b T})$	$N_c^{2d} F_0(\frac{F_n - E_c}{k_b T})$	$N_c^{3d} F_{+\frac{1}{2}}(\frac{F_n - E_c}{k_b T})$	$N_c^{d} F_{\frac{d-2}{2}}(\frac{F_n - E_c}{k_b T})$
Hole Density in the Optically Active Region	–	$N_v^{1d} F_{-\frac{1}{2}}(\frac{E_v - F_p}{k_b T})$	$N_v^{2d} F_0(\frac{E_v - F_p}{k_b T})$	$N_v^{3d} F_{+\frac{1}{2}}(\frac{E_v - F_p}{k_b T})$	$N_v^{d} F_{\frac{d-2}{2}}(\frac{E_v - F_p}{k_b T})$
Excess Electron-Hole Pair Density	–	$N_c^{1d} N_v^{1d}[F_{-\frac{1}{2}}(\frac{F_n - E_c}{k_b T}) F_{-\frac{1}{2}}(\frac{E_v - F_p}{k_b T}) - F_{-\frac{1}{2}}^2(\frac{E_F - E_c}{k_b T})]$	$N_c^{2d} N_v^{2d}[F_0(\frac{F_n - E_c}{k_b T}) F_0(\frac{E_v - F_p}{k_b T}) - F_0^2(\frac{E_F - E_c}{k_b T})]$	$N_c^{3d} N_v^{3d}[F_{+\frac{1}{2}}(\frac{F_n - E_c}{k_b T}) F_{+\frac{1}{2}}(\frac{E_v - F_p}{k_b T}) - F_{+\frac{1}{2}}^2(\frac{E_F - E_c}{k_b T})]$	$N_c^{d} N_v^{d}[F_{\frac{d-2}{2}}(\frac{F_n - E_c}{k_b T}) F_{\frac{d-2}{2}}(\frac{E_v - F_p}{k_b T}) - F_{\frac{d-2}{2}}^2(\frac{E_F - E_c}{k_b T})]$
Spontaneous Emission Spectrum $R_{sp}(\nu)$	–	$A \cdot \frac{g_s g_v}{2\pi L_y L_z}(\frac{2m_r^\star}{\hbar^2})^{\frac{1}{2}} \frac{1}{\sqrt{h\nu-E_g}} \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$	$A \cdot \frac{g_s g_v}{2\pi L_z}(\frac{2m_r^\star}{\hbar^2})\theta(h\nu - E_g) \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$	$A \cdot \frac{g_s g_v}{(2\pi)^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{h\nu - E_g} \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$	$A \cdot \rho_{jnt}(\nu) \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$
Stimulated Emission Spectrum $R_{st}(\nu)$	–	$B \cdot \rho_\nu \cdot \frac{g_s g_v}{2\pi L_y L_z}(\frac{2m_r^\star}{\hbar^2})^{\frac{1}{2}} \frac{1}{\sqrt{h\nu-E_g}} \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$	$B \cdot \rho_\nu \cdot \frac{g_s g_v}{2\pi L_z}(\frac{2m_r^\star}{\hbar^2})\theta(h\nu - E_g) \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$	$B \cdot \rho_\nu \cdot \frac{g_s g_v}{(2\pi)^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{h\nu - E_g} \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$	$B \cdot \rho_\nu \cdot \rho_{jnt}(\nu) \cdot f_c(E_2) \cdot [1 - f_v(E_1)]$
Absorption Spectrum $R_{ab}(\nu)$	–	$B \cdot \rho_\nu \cdot \frac{g_s g_v}{2\pi L_y L_z}(\frac{2m_r^\star}{\hbar^2})^{\frac{1}{2}} \frac{1}{\sqrt{h\nu-E_g}} \cdot f_v(E_1) \cdot [1 - f_c(E_2)]$	$B \cdot \rho_\nu \cdot \frac{g_s g_v}{2\pi L_z}(\frac{2m_r^\star}{\hbar^2})\theta(h\nu - E_g) \cdot f_v(E_1) \cdot [1 - f_c(E_2)]$	$B \cdot \rho_\nu \cdot \frac{g_s g_v}{(2\pi)^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{h\nu - E_g} \cdot f_v(E_1) \cdot [1 - f_c(E_2)]$	$B \cdot \rho_\nu \cdot \rho_{jnt}(\nu) \cdot f_v(E_1) \cdot [1 - f_c(E_2)]$
Photonic Gain Spectrum of Semiconductor $\gamma_0(\nu)$	–	$A \cdot \frac{\lambda_0^2}{8\pi n^2} \cdot h \frac{g_s g_v}{2\pi L_y L_z}(\frac{2m_r^\star}{\hbar^2})^{\frac{1}{2}} \frac{1}{\sqrt{h\nu-E_g}} \cdot [f_c(E_2) - f_v(E_1)]$	$A \cdot \frac{\lambda_0^2}{8\pi n^2} \cdot h \frac{g_s g_v}{2\pi L_z}(\frac{2m_r^\star}{\hbar^2})\theta(h\nu - E_g) \cdot [f_c(E_2) - f_v(E_1)]$	$A \cdot \frac{\lambda_0^2}{8\pi n^2} \cdot h \frac{g_s g_v}{(2\pi)^2}(\frac{2m_r^\star}{\hbar^2})^{\frac{3}{2}} \sqrt{h\nu - E_g} \cdot [f_c(E_2) - f_v(E_1)]$	$A \cdot \frac{\lambda_0^2}{8\pi n^2} \cdot h\rho_{jnt}(\nu) \cdot [f_c(E_2) - f_v(E_1)]$

Table 1: Quantum Electronic, Photonic, and Statistical Properties of Conduction and Valence Band Electrons in Semiconductor Nanostructures.

- E_c is the band edge, and $m_c^\star$ the effective mass of the conduction band. E_v is the band edge, and $m_v^\star$ the effective mass of the valence band.
- For low dimensions, E_c and E_v, and the bandgap $E_g = E_c - E_v$ include the quantum confinement energies if present.
- h is Planck's constant, $\hbar = \frac{h}{2\pi}$, k_b is the Boltzmann constant, and q the electron charge.
- g_s is the spin degeneracy, and g_v the valley degeneracy.
- $F_j(\eta) = \frac{1}{\Gamma(j+1)} \int_0^\infty du \frac{1}{1+e^{u-\eta}}$ is the Fermi-Dirac integral of order j, and $\Gamma(...)$ is the Gamma function.
- E_F is the Fermi level at equilibrium. E_{Fs} is the source quasi-Fermi level and E_{Fd} the drain quasi-Fermi level.
- Similarly, F_n is the conduction band quasi-Fermi level and F_p is the valence band quasi-Fermi level.
- For interband tunneling, F is the junction field, $F_0 = \frac{4}{3}\sqrt{\frac{2m_r^\star E_g^3}{\hbar^2 q^2}}$ is the characteristic tunneling electric field, and $qV_0 = \eta^2 \bar{E}$, where $\bar{E} = \frac{2F}{3F_0} E_g$ and $\eta^2 = m_v^\star / m_r^\star$.
- $h\nu$ is the photon energy of frequency ν, and L_x, L_y, L_z are the dimensions of the semiconductor nanostructure.
- A and B are the Einstein A and B coefficients, $\lambda_0 = c/\nu$ the wavelength of the photon in vacuum, and n the refractive index of the semiconductor.
- $f_c(E_2) = 1/[1 + \exp[(E_2 - F_n)/(k_b T)]]$ is the Fermi-Dirac occupation function of state $E_2 = E_c + \frac{\hbar^2 k^2}{2m_c^\star}$ in the conduction band.
- $f_v(E_1) = 1/[1 + \exp[(E_1 - F_p)/(k_b T)]]$ is the Fermi-Dirac occupation function of state $E_1 = E_v - \frac{\hbar^2 k^2}{2m_v^\star}$ in the valence band.
- $E_2 - E_1 = h\nu = E_g + \frac{\hbar^2 k^2}{2m_r^\star}$ is the energy of the photon emitted when the electron transitions from $E_2 \to E_1$ radiatively.

Table A.1 Physical properties of selected 3D semiconductors and insulators.

Property	Symbol	Units	Si	Ge	GaAs	InP	SiC (6H)	GaN	SiO2	Si_3N_4	HfO2
Crystal Structure			Diamond	Diamond	Zincblende	Zincblende	Wurtzite	Wurtzite	Amorphous		Amorphous
Lattice Constant	a (c)	Å	5.4307	5.6575	5.6532	5.8686	3.0730(10.053)	3.191(5.185)		7.75	
Density		g/cm^3	2.328	5.323	5.316	4.81	3.21	6.15	2.19	3.44	9.68
Direct/Indirect			Indirect	Indirect	Direct	Direct	Indirect	Direct			
Energy Bandgap (300K)	E_g	eV	1.124	0.67	1.42	1.35	3.03	3.425	9	4.7	5.8
Energy Bandgap (0K)	E_g	eV	1.17	0.744	1.52	1.421					
Energy Bandgap T-dependence	dE_g/dT	10^{-4} eV/K	-2.7	-3.7	-5						
Relative Dielectric Constant	ϵ_r		11.7	16	13.1	12.4	9.66(10.33)	8.9	3.9	7.5	25
Refractive index	n_r		3.44	3.97	3.33	3.4	2.55	2.5	1.46	2	2.1
Melting Point	T_m	°C	1412	937	1237	1060	2830	1600	1700	1900	2758
Thermal Conductivity	κ	W/m.K	141.2	60.6	45.5	68	490	240	1.4	18.5	
Thermal Expansion Coefficient	$\alpha(a, c)$	10^{-6} /K	2.5	5.7	5.9	4.6	4.3(4.7)	5.59(3.17)	0.5	2.8	
Effective mass: Electrons	$m_c^\star$ (m_l, m_t)	m_e	0.26	0.12	0.068	0.08	0.2(0.42)	0.2			
Effective mass: Holes	$m_v^\star$	m_e	0.386	0.3	0.5	0.56		0.8			
Intrinsic Concentration	n_i	cm^{-3}	1.45×10^{10}	2.4×10^{13}	9×10^6	1×10^7	~ 0	~ 0			
Band-edge DOS: CB	N_c	cm^{-3}	2.8×10^{19}	1.04×10^{19}	4.7×10^{17}	5.2×10^{17}		2.3×10^{18}			
Band-edge DOS: VB	N_v	cm^{-3}	1.04×10^{19}	6.0×10^{18}	7×10^{18}	1.1×10^{19}		1.8×10^{19}			
Mobility: Electrons	μ_n	$cm^2/V \cdot s$	1417	3900	8800	4600	$400 \perp c, 60 \parallel c)$	1500	20		
Mobility: Holes	μ_p	$cm^2/V \cdot s$	471	1900	400	150	90	30	0		
Critical breakdown field	F_{cr}	10^6 V/cm	0.3	0.08	0.35	0.5	$3.2(\parallel c)$	3.5	9		
Electron Affinity	$q\chi$	eV	4.05	4	4.07	4.5		4.1	1		
Optical Phonon Energy	$\hbar\omega_{op}$	meV	63	37	35	42.6	104.2	92			
References			R9	R9	R9	R9	R1, R41	R1, R11	R9	R9	R40

Table A.2 Extended table of properties of 3D semiconductors

	Semiconductor	Crystal Structure	Bandgap eV	Lattice Constant a Å	b Å	c Å	Dielectric Constant ϵ_r	ϵ_∞	Refractive Index n_r	Optical Phonon $\hbar\omega_{op}$ meV	Cond Band Effective mass $\frac{m_l}{m_e}$	m_t	Valence Band Effective mass $\frac{m_{lh}}{m_e}$	m_{hh}	Deformation Potential D_c eV	D_v	Mass Density ρ g/cm^3	Sound Velocity $v_{s,l}$ 1×10^5 cm/s	$v_{s,t}$	Piezoelectric Coefficient e_{14} or e_{33} C/m^2	e_{31}	e_{15}	Thermal Conductivity κ W/m.K	References
3D Crystals Group IV	C	Diamond	5.49	3.56683			5.7		2.4	160	1.4	0.36	0.7	2.12			3.515	17.6	12.8				2000	R1, R
	SiC (6H)	Diamond	3.02	3.0805	15.1151		9.66(10.03)	6.52	2.55	104.2	0.2	0.42					3.21	12.5	7.1				490	R1, R3, R
	Si	Diamond	1.124	5.4307			11.7		3.44	63	0.98	0.19	0.16	0.49	9	5	2.329	8.432	5.842				141.2	R1, R3, R5
	Ge	Diamond	0.67	5.6575			16		3.97	37	1.6	0.08	0.043	0.33	11.5		5.3234	4.914	3.543				60.6	A1, A3, A
III-Nitrides	BN	Cubic/Graphitic	5.955					7.1	4.5	130							2.18						740	R1, R3, R
	AlN	WZ	6.28Γ	3.112		4.982	8.5	4.84	2.15	99.2	0.4	0.4	0.24	10.42	9.5		3.23	11.27	6.22	0.97	0.57		340	R1, R3, R4
	GaN	WZ	3.425Γ	3.191		5.185	8.9	5.5	2.5	91.1	0.2	0.2	0.3	1.4	8.3	6.2	6.1	7.96	5.02	0.65	0.33	0.3	260	R1, R3, R4
	InN	WZ	0.65Γ	3.545		5.703	15.3	8.84	2.9	73	0.11	0.11	0.27	1.63	7.1		6.81	5.25	1.21	1.55	0.58	0.48	176	R1, R3, R4
									2.9															
III-Phosphides	AlP	ZB	2.43X	5.4625			9.8	7.5	3															R1, R3, R
	GaP	ZB	2.26X	5.4506			11.1	9.11	3.37	51	1.12	0.22	0.14	0.79	13.0		4.14	5.83	4.12	−0.1			110	R1, R3, R
	InP	ZB	1.35Γ	5.8688			12.6	9.61	3.37	43	0.08		0.089	0.6	6.8		4.81	4.58	3.08	−0.035			68	R1, R3, R
III-Arsenides	AlAs	ZB	2.16X	5.6611			10.1		3.2	50.09			0.49	1.06			3.76	5.65	3.96				84	R1, R3, R
	GaAs	ZB	1.42Γ	5.6533			13.1		3.4	50.09	0.063		0.12	0.51	6.3	3.5	5.32	4.73	3.35	−0.16			45.5	R1, R3, R
	InAs	ZB	0.36Γ	6.0584			15.1		3.5	50.09	0.023		0.026	0.41	5.8		5.68	3.83	2.64	−0.045			27	R1, R3, R
III-Antimonides	AlSb	ZB	1.52X	6.1355			11		3.4								4.26							R3
	GaSb	ZB	0.72Γ	6.0954			15.69	14.44	3.9	29.7	0.041		0.05	0.4	8.3		5.61	3.97	2.77	−0.13			32	R3
	InSb	ZB	0.18Γ	6.4788			17.64	15.75	3.75	25	0.014		0.015	0.43	7.2		5.77	3.4	2.29	−0.07			18	R3
II-Oxides	ZnO	WZ	3.3Γ	3.2501		5.2069	8.87	3.75	2.01	72	0.24						5.66			1.32	−0.57			R3
	CdO	Cubic					18.1	5.4	2.49								8.15							R2, R
	HgO								2.36								11.14							R3
II-Sulfides	ZnS	ZB	3.6Γ	3.8227			9.6	5.4	2.4		0.39				4.9		4.09							R3
	CdS	WZ	2.42Γ	4.136			8.53	5.26	2.5		0.2				3.3		4.82							R3
	HgS	ZB	2.0Γ	5.852			10.33	8.4	2.55								8.1							R3, R
II-Selenides	ZnSe	ZB	2.58Γ	5.6686			8.6	5.7	2.89		0.17				4.2		5.27							R3, R
	CdSe	WZ	1.73Γ	6.078			8.6	6.1	2.5		0.2				3.7		5.82							R3, R
	HgSe	ZB	−0.15Γ	6.084			25.6	12							4.0		8.3							R3, R
II-Tellurides	ZnTe	ZB	2.25Γ	6.1004			9.1	6.75	3.56		0.15				3.5		6.34							R3, R
	CdTe	ZB	1.50Γ	6.481			10.2	7.25	2.75		0.11		0.35		4.0		5.85							R3, R
	HgTe	ZB	−0.15Γ	6.462			20	14	3.7		0.029		0.35		4.0		8.1							R3, R
													0.3											
Other Oxides	α–Al$_2$O$_3$	Corundum	8.7	4.81		13.12	10.07	3.1	1.76		0.44	0.44					3.99						35	R17
	α–Ga$_2$O$_3$	Corundum	5.2	4.98		13.43		3.03	1.74		0.32(m_z)	0.30(m_x)					6.47							R18, R
	α–In$_2$O$_3$	Corundum																						
	β–Al$_2$O$_3$	Monoclinic																						
	β–Ga$_2$O$_3$	Monoclinic	4.9	12.23	3.06	5.81	10.02	3.56	1.89	44	0.28	0.28			6.9		5.88	6.8	3.4				16.3	R12-R
	β–In$_2$O$_3$	Monoclinic																						
	In$_2$O$_3$	Cubic/Bixbyite	2.9	10.12			10.55	4.05	2.01		0.21	0.21					7.18							R17, R20
	BaSnO$_3$	Perovskite	3.1	4.12																				R42
	SrTiO$_3$	Perovskite	3.2	3.905			310	5.2	2.28	11	1.41	1.41											110	R23-R
Lead Chalcogens	PbS	NaCl	0.37L	5.936			175	17	3.7		0.1		0.1		20		7.6						250	R3, R8, R
	PbSe	NaCl	0.26L	6.147			250	24			0.07	0.039	0.06	0.03	24		8.15						162	R3, R
	PbTe	NaCl	0.29L	6.45			400	33			0.24	0.02	0.3	0.02	25		8.16						199	R3, R8, R

Table A.3 Table of properties of 2D semiconductors

	E_g eV	$\frac{m^\star}{m_e}$	D_a eV	ρ $10^{-6}\,\frac{\text{kg}}{\text{m}^2}$	v_s $10^3\,\frac{\text{m}}{\text{s}}$	$e_{pz}(e_{11})$ $10^{-11}\,\frac{\text{C}}{\text{m}}$	$\hbar\omega_{op}$ meV	D_{op} $\frac{\text{eV}}{\text{nm}}$	$\frac{\epsilon_{dc}}{\epsilon_0}$	$\frac{\epsilon_\infty}{\epsilon_0}$	t_0 Å	$\frac{\text{atoms}}{\text{cell}}$	References
			LA TA		L T		$\hbar\omega_{LO}$ $\hbar\omega_{HP}$	LO HP	$\perp$ $\parallel$	$\perp$ $\parallel$			
C (graphene)	0		5.1 2.8	0.76	21 14		196	79				2	R27-R29
BN (monolayer)	> 5 eV					1.38			6.82 3.29	4.97 2.89	3.17	2	R30, R31
P (Phosphorene)	1.75	$x : 0.18; y : 1.23$			$x : 4.1; y : 7.8$ $x : 3.6; y : 3.3$		54					4	R32
MoS_2	1.72	0.56	2.4 1.5	3.1	6.7 4.2	3.64	48 50	26 41	6.2 15.5	6.1 15.4	5.41	3 (1H)	R30, R31, R33
WS_2	1.93	0.33			6.2 4.0	2.47	44 52		6.1 14.0	6.1 14.0	6.12	3 (1H)	R30, R31, R33
ZrS_2	1.51								6.8 66.7	5.7 11.2	5.74	3 (1H)	R30, R31, R33
HfS_2	2.02								6.6 53.6	5.6 10.2	5.74	3 (1H)	R30, R31, R36
$MoSe_2$	1.49	0.59			5.9 3.4	3.92	34 31		7.2 17.4	7.2 16.8	6.49	3 (1H)	R30, R31, R33, R37
WSe_2	1.62	0.33			4.8 3.0	2.71	31 32		7.4 15.6	7.4 15.3	6.48	3 (1H)	R30, R31, R38
$HfSe_2$	1.07								7.6 83.0	6.7 13.9	6.13	3 (1H)	R30, R31, R36
$MoTe_2$	1.13	0.60				5.43			9.1 21.6	8.9 19.7	7.1	3 (1H)	R30, R31, R33, R39

A.4 References for the appendix

- R1 `http://www.ioffe.ru/`
- R2 `https://en.wikipedia.org/wiki/Cadmium_oxide`
- R3 Charles M. Wolfe, Nick Holonyak, Jr., Gregory E. Stillman, Physical properties of semiconductors
- R4 Karlheinz Seeger, Semiconductor Physics
- R5 Cardona, M., Kremer, R.K., Siegle, G., Muoz, A., Romero, A.H., Schmidt, M.: Phys. Rev. B 82 (2010) 085210.
- R6 Popov, A. S., Trifonova, E. P.: Phys. Status Solidi (a) 58 (1980) 679
- R7 Kug Sun Hong, Speyer, R.F., Condrate, R.A.: J. Phys. Chem. Solids 51 (1990) 969.
- R8 A. A. El-Sharkawy, A. M. Abou EI-Azm, M. I. Kenawy, A. S. Hillal, and H. M. Abu-Basha, I. J. Thermo. 4, 261 (1983)
- R9 Richard Muller, Theodore Kamins, and Mansun Chan, Device Electronics for Integrated Circuits
- R10 Chihiro Hamaguchi, Basic Semiconductor Physics
- R11 E. Fred Schubert, Light-emitting Diode
- R12 Hoeneisen, B., Mead, C. A. and Nicolet, M. A.: Solid. State. Electron. 14, 10571059 (1971).
- R13 Sturm, C., Furthmüller, J., Bechstedt, F., Schmidt-Grund, R. and Grundmann, M.: APL Mater. 3, 09 (2015).
- R14 Parisini, A. and Fornari, R.: Semicond. Sci. Technol. 31, (2016).
- R15 Guo, Z. et al.: Appl. Phys. Lett. 106, 16 (2015).
- R16 Ma, N. et al.: Appl. Phys. Lett. 109, 16 (2016).
- R17 CRC handbook of chemistry and physics. CRC Press.
- R18 He, H. et al.: Phys. Rev. B - Condens. Matter Mater. Phys. 74, 18 (2006).
- R19 Feneberg, M. et al.: Appl. Phys. Lett. 114, (2019).
- R20 Marezio, M.: Acta Crystallogr. 20, 723728 (1966).
- R21 King, P. D. C. et al.: Phys. Rev. B - Condens. Matter Mater. Phys. 79, 110 (2009).
- R22 Stokey, M. et al.: J. Appl. Phys. 129, (2021).
- R23 Neville, R. C., Hoeneisen, B. and Mead, C. A.: J. Appl. Phys. 43, 21242131 (1972).
- R24 Allen, S. J. et al.: Phys. Rev. B 88, 045114 (2013).
- R25 Yamada, Y. and Shirane, G.: J. Phys. Soc. Japan 26, 396403 (1969).
- R26 Suemune, Y.: J. Phys. Soc. Japan 20, 174175 (1965).
- R27 Castro, E. V. et al.: Phys. Rev. Lett. 105, 1618 (2010).

- R28 Kaasbjerg, K., Thygesen, K. S. and Jacobsen, K. W.: Phys. Rev. B - Condens. Matter Mater. Phys. 85, 79 (2012).
- R29 Lazzeri, M., Attaccalite, C., Wirtz, L. and Mauri, F.: Phys. Rev. B - Condens. Matter Mater. Phys. 78, 081406 (2008).
- R30 Laturia, A., Van de Put, M. L. and Vandenberghe, W. G.:. npj 2D Mater. Appl. 2, (2018).
- R31 Duerloo, K. A. N., Ong, M. T. and Reed, E. J.: J. Phys. Chem. Lett. 3, 28712876 (2012).
- R32 Carvalho, A. et al.: Nat. Rev. Mater. 1, (2016).
- R33 Liu, L., Kumar, S. B., Ouyang, Y. and Guo, J.: IEEE Trans. Electron Devices 58, 30423047 (2011).
- R34 Molina-Snchez, A. and Wirtz, L.: Phys. Rev. B - Condens. Matter Mater. Phys. 84, 18 (2011).
- R35 Li, Y., Kang, J. and Li, J.: RSC Adv. 4, 73967401 (2014).
- R36 Zhao, Q. et al.: Phys. Status Solidi Basic Res. 254, (2017).
- R37 Horzum, S. et al.: Phys. Rev. B - Condens. Matter Mater. Phys. 87, 15 (2013).
- R38 Sengupta, A., Chanana, A. and Mahapatra, S.: AIP Adv. 5, (2015).
- R39 Ruppert, C., Aslan, O. B. and Heinz, T. F.: Nano Lett. 14, 62316236 (2014).
- R40 Robertson, Eur. Phys. J. Appl. Phys. 28, 265291 (2004)
- R41 `https://www.iue.tuwien.ac.at/phd/ayalew/node21.html`
- R42 Kim, H. J. et al., Appl. Phys. Exp. 5, 061102 1-3 (2012)

A.5 Physical constants

See the next page.

Table A.4 Table of physical constants

Speed of light in vacuum	c	2.998×10^{8}	m/s
Permittivity of free space	ϵ_0	8.854×10^{-12}	F/m
Permeability of free space	μ_0	$4\pi \times 10^{-7}$	H/m
Electron charge	$-q$	-1.602×10^{-19}	C
Free electron mass	m_e	9.109×10^{-31}	kg
Proton charge	$+q$	1.602×10^{-19}	C
Proton mass	m_p	1.673×10^{-27}	kg
Neutron mass	m_n	1.674×10^{-27}	kg
Boltzmann's constant	k_b	1.381×10^{-23}	J/K
Planck's constant	h	6.626×10^{-34}	J.s
Reduced Planck's constant	$\hbar = \frac{h}{2\pi}$	1.054×10^{-34}	J.s
Bohr radius	a_B^0	5.292×10^{-11}	m
Rydberg energy	Ry^0	13.61	eV
Bohr magneton	μ_B	9.274×10^{-24}	J/T
Magnetic flux quantum	ϕ_0	4.136×10^{-15}	Wb
Conductance quantum	$G_0 = \frac{2q^2}{h}$	7.748×10^{-5}	S